ISSN 1331-7210

1993. 2018.

Engineering Power

BULLETIN OF THE CROATIAN ACADEMY OF ENGINEERING Vol. 13(2) 2018

Euro-CASE CAETS

EDITOR-IN-CHIEF'S WORD

Dear readers,

In its determination to bring you closer to the activities and works of its members the Croatian Academy of Engineering is traditionally publishing our HATZ Bulletin *Engineering Power (Vol. 13/2018)* whose guest editor as a member of our Academy is presenting his achievements in his field of expertise.

As the Academy in its commitment insists on *multidisciplinarity* without which it is difficult to imagine inclusion in the already existing level of Industry 4.0 we have asked our distinguished member of the Croatian Academy of Engineering, Department of Systems and Cybernetics and a distinguished Associate Professor at the Faculty of Electrical Engineering and Computing of the University of Zagreb, that he as guest editor presents some of his achievements in a computer approach to some areas of artificial intelligence.

I believe that you will find appropriate interest and new knowledge in this field.

Editor-in-Chief
Vladimir Andročec, President of the Croatian Academy of Engineering

EDITOR'S WORD

This issue of the HATZ Bulletin Engineering Power continues to present multidisciplinary research activities of the Academy members that are actual and have wide application areas, and thus they synthesise a number of advanced scientific disciplines and exert significant influence on numerous modern living areas. Guest Editor is Tomislav Pribanić, Ph.D., Associate Professor, Faculty of Electrical Engineering and Computing, University of Zagreb, associate member of the Academy and Head of one of the laboratories of the Center for Excellence for Computer Vision (CRV) at the same Faculty.

Editor
Zdravko Terze, Vice-President of the Croatian Academy of Engineering

FOREWORD

A large part of present technological achievements results from research and continuing advances in the field of artificial intelligence (AI). AI is a part of computer science that aims to create intelligent machines, capable of thinking, acting and learning like humans. It is an interdisciplinary field spanning a variety of subfields, among which machine learning (ML) and computer vision (CV) are generally regarded as core parts of AI. CV is a field that aims to give the computer visual understanding of the world from images. ML is a field of study that gives computers the ability to learn how to solve a certain task. It is particularly suited for problems that may seem relatively simple for humans, but are rather difficult to solve by using classical image processing approaches. CV and ML fields have a significant overlap where many CV problems can be solved using ML techniques.

Several papers listed below present a part of CV and related ML research conducted by experts from two laboratories of the Center of Excellence for Computer Vision (CRV) at the University of Zagreb Faculty of Electrical Engineering and Computing and also by CRV collaboration researchers. The first laboratory involved is Human-oriented Technologies Laboratory (HOTLab) led by Prof. Igor S. Pandžić, Ph.D., while the second laboratory involved is Advanced Shape Reconstruction and Registration Laboratory (SHARK Lab) led by Tomislav Pribanić, Ph.D., Associate Prof.

Nowadays one heavily studied ML application is certainly face analysis (FA) presented in one of the papers below. Applications of FA technologies range from marketing and entertainment to automotive industry in which, for instance, the goal is fatigue detection for vehicle driver. Another paper presented discusses two thoroughly researched CV tasks: object localization and semantic segmentation. The former attempts to find objects in the input image, where minimum bounding rectangle of the object and the associated object class are the ideal output. The latter is somewhat more detailed where each image pixel is assigned to the corresponding class label. Interesting applications can be found in traffic control systems and medical imaging. The next paper presents ML in the context of image categorization and image similarity whereby a commercial service was developed, enabling buyers of certain products to find visually similar objects of interest. The camera is the essential tool used in CV. For numerous geometry related tasks the camera requires calibration which affects many applications such as geocoding, as explained in another paper. A geometrically calibrated camera is a basis for the 3D passive and 3D active reconstruction system too. 3D scanning systems are extensively used in fashion design and development and medical applications such as human back surface analyses. The last two papers put emphasis on those two applications.

Guest-Editor
Tomislav Pribanić, University of Zagreb Faculty of Electrical Engineering and Computing

CONTENT

Krešimir Bešenić[1], Ivan Gogić[2], Igor S. Pandžić[2] and Krešimir Matković[3]

Automatic Image-based Face Analysis Systems Overview

[1]Visage Technologies, Diskettgatan 11A, SE-583 35 Linköping, Sweden
[2]University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
[3]VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, Donau-City-Straße 11, 1220 Wien, Austria

Abstract

Face analysis systems have recently gained popularity due to the large number of potential applications across a wide range of industries. Various types of information can be extracted from an image of a face including: face location and size, location of characteristic facial landmark points, 3D head pose, facial expression and emotion, gaze direction and biometric information (i.e. age, gender and race). Most of these problems are solved using machine learning techniques based on large sets of training samples. Furthermore, information from these different tasks is often complementary and can be used to enhance the accuracy of the algorithms. A systematic overview of current approaches to face analysis tasks is presented as an introduction to this growing research field.

1. Introduction

Applications of Face Analysis (FA) technologies span numerous and diverse industrial and commercial fields.

Currently the most common applications are found in **marketing and entertainment**, based on the novelty and fun effect of FA, usually combined with 3D graphics, such as the popular face masks in Snapchat and similar apps. Numerous major brands have used FA effects in their online marketing campaigns. Furthermore, products such as make-up, glasses or even hats and earrings, use **virtual try-on** applications for promotion and testing. Many such applications allow direct purchasing. In physical **retail** spaces, experiments are starting to analyze customer behavior and shopping patterns using cameras placed in shops or shop windows. In **marketing research**, analysis of subjects' gaze patterns and emotional reactions has traditionally been performed in on-site studies using specialized gaze tracking hardware and requiring large number of subjects. The new generation of marketing research technology uses FA software to perform similar research with subjects participating from home (being paid as micro-workers), dramatically reducing cost and increasing speed and scale of possible research. **Automotive** industry has been deploying various forms of fatigue detection in heavy commercial vehicles and, more recently, in cars. However, the use of FA for driver monitoring is still in fairly early stages and we expect to see much more widespread deployment in years to come. Furthermore, there is interest in other uses of FA such as controlling the information system or automatic personal adjustments in high-end cars. By monitoring operators of various types of machinery (e.g. forklifts), FA can help increase **industrial safety**. **Assistive technologies** help people with disabilities perform various tasks by using limited movement such as gaze or head motion. **Biometrics** based on face recognition is increasingly deployed for access control (e.g. to financial services). To avoid trivial fraud by submitting a picture instead of the live face, such applications deploy liveness detection techniques based on FA. Ubiquitous computing power and variety of available sensors are already changing the way we treat **health**, allowing simplified and more widespread monitoring and diagnostics through inexpensive devices and apps. FA technologies play a role in this trend, with experimental or prototype applications for remote fever detection, posture monitoring, concussion diagnostics and others. Further applications of FA include **robotics**, where it allows robots to interact with humans, and advanced **audio** systems that use 3D head position to deliver perfect sound to the listener.

2. General Face Analysis Framework

A typical face analysis framework can be viewed as a pipeline consisting of several steps. As in many other image analysis frameworks, the first step is object (face) detection. The face detection step is usually followed by preprocessing, face alignment, feature extraction, and attribute prediction steps, sequentially (Figure 1).

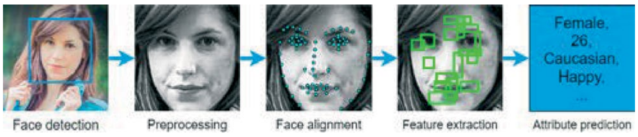


Fig. 1. General face analysis framework

2.1. Face Detection

In the proposed framework pipeline, initial face detection step is semi-decoupled from other steps as it results in basic face location and scale information, usually provided via a facial bounding box. Even though the bounding box information is most basic, different versions of face detectors can be trained with differently defined bounding boxes and can results in different detection qualities, thus introducing bias and propagating the error to the rest of the pipeline. To some extent, this can be alleviated in a preprocessing step and by introducing perturbation augmentations to the training set. Most widely used face detection systems are based on the work of Viola and Jones [1] and, more recently, deformable parts models [2] and single shot detection systems [3]. More details on face detection methods can be found in [4]

2.2. Preprocessing

Depending on the face analysis method, the preprocessing step can be as trivial as image cropping based on the facial bounding box. Typical minimal preprocessing techniques also include resizing and color conversions. To deal with low contrast and lighting problems, additional preprocessing techniques such as histogram equalization, Difference of Gaussians filtering, and edge enhancement filtering (e.g. Sobel filtering) can be incorporated.

2.3. Face Alignment

To compensate for face detector inaccuracies and to deal with misaligned faces captured in unconstrained conditions, various alignment techniques have been proposed. Most basic method rests on face detection confidence. The input image is rotated by a small angle multiple times and the version with the highest detection confidence is used. Although computationally expensive, this simple method does not introduce any new components (existing face detection system is reused) and can result in a satisfying performance. More complex methods rely on facial landmark point detection methods. Given a set of detected landmark points, in-plane rotation and scaling can be performed based on eye points locations, Procrustes Analysis transform or 3D face model fitting.

2.4. Feature extraction

It is well known that any classifier is only as good as the data it works with. This applies to all types of face analysis systems, therefore in many cases making the feature extraction step the most important one. Geometric feature extraction, which is based on fiducial distance measurements, heavily relies on precise facial landmark points detection. Geometric features can be reliable in 3D use-cases, yet in the case of 2D images, their practicality is usually restricted to constrained frontal neutral faces. Appearance-based features consist of pixel values or their transformations, thus making them more suitable for 2D image use. While raw pixel values can be used directly as an input for classification and regression systems, more elaborate approaches such as Local Binary Patterns (LBP), Biologically Inspired Features (BIF), Haar-like features, Histogram of Oriented Gradients (HOG) features, speeded up robust features (SURF), and Gabor filters are commonly used.

2.5. Attribute prediction

In this work, attribute prediction refers to classification and regression tasks related to expression, age, gender, and race prediction. Binary classification is commonly used for gender classification and simple race classifiers (e.g. Asian/Non-Asian, White/Black). Multi-class classification is used for face expression and age group classifiers, and in some cases for exact age prediction (e.g. 100 classes, one for each year). Regression is a natural (but not necessarily optimal) choice for exact age estimation. Han et al. combined classification and regression in their proposed hierarchical estimator consisting of a between-group classification and a within-group regression in [5].

Some methods, most notably neural networks, perform multiple steps in a joint manner. Deep Convolutional Neural Networks (DCNN) combine feature extraction and attribute prediction steps together to learn optimal feature extractors and model high-level abstractions from the data. Due to their flexibility, DCNNs can be used for classification, regression or for more elaborate combinations of those two approaches.

3. Face Alignment

Face alignment is the process of determining the location of characteristic facial features or landmarks (points that delineate eyes, nose, mouth, eyebrows, chin and face contour) given a face image. The configuration of facial landmarks is also usually referred to as face shape which is represented as a vector of 2D landmark coordinates. Various machine learning algorithms are employed in order to estimate the face shape. If we denote it with $S = (x_1, y_1,..., x_L, y_L)$ where L represents the number of landmarks, the goal of face alignment, given a face image, is to find a shape S closest to the ground truth shape S^*. More formally, the goal is to minimize:

$$\|S - S^*\| \tag{1}$$

where $\|\cdot\|$ is a suitable vector norm. The alignment error in (1) is used as a performance measure that drives the training process.

Regression methods estimate the face shape directly from image features and have recently demonstrated superior accuracy, speed and robustness when compared to earlier, traditional methods that involve Active Appearance Models, Active Shape Models and local part classification using search algorithms. Such constructed models demonstrate poor ability to express all combinations of face variations due to expressions, illumination and head pose [6].

Regression methods can be roughly divided into four categories: constrained regression, cascaded regression, deep learning, and head pose and occlusion methods. Constrained regression methods estimate landmark positions individually, then additionally ensure a probable face configuration. However, in cascaded regression framework, an implicit face shape constraint is incorporated into the training process. This framework is currently the standard approach to face alignment. Recently, with advances in computing power and optimization techniques, Convolutional Neural Networks (CNN) have been applied to face alignment as part of the deep learning category. With the growing success of cascaded regression and deep learning methods, face alignment in more challenging conditions has become the focus area for researchers as part of the head pose and occlusion category.

Cascaded regression has established itself as the leading approach for face alignment due to its speed, robustness and accuracy. In this framework, a number of regressors $(R_1 ,...,R_t ,...,R_T)$ are successively applied starting from the initial shape estimate S_0. Given an image I, each regressor learns and estimates a shape increment δS and updates the face shape:

$$\delta \mathrm{S} = \mathrm{R}_t\, (I,S_{t-1}) \tag{2}$$

$$S_t = S_{t-1} + \delta S \tag{3}$$

where the tth regressor R_t updates the previous shape S_{t-1} to the new shape S_t (Cao et al. 2014). It is important to note that the tth regressor depends on the previous shape estimate S_{t-1}. The dependency is usually through shape-indexed features which is a concept first introduced in [7]. These features are stored relative to the object pose and are thus consistent across large pose variations.

4. Facial Expression Recognition

In order to automatically recognize emotions and their related expressions, an investigation on how to define those terms needed to be done first. In [8], Ekman and Friesen discovered six basic or prototypic emotions (anger, disgust, fear, happiness, sadness, and surprise) whose facial expressions are culturally and racially invariant and are, therefore, great candidates for automatic systems which need clear categories. However, one important drawback of this model became evident. It is too crude to accurately model the complexity of emotions people experience in everyday lives. As a response, Facial Action Coding System (FACS) was developed in order to define atomic facial muscle movements named Action Units (AU) spanning the whole spectrum of human facial expressions. Its aim is objectivity in the signal measurement which is separated from the final expression classification often influenced by the context. Consequentially, a group of researchers tried to develop algorithms that recognize these simpler, intermediate categories and synthesize the final expression afterward. However, FACS annotation is a very tedious process which requires expert knowledge few people poses. Therefore, few data sets with full FACS annotations are available to the community. The six basic expressions classification approach is currently the most widely used categorization in computer vision.

The features used for Facial Expression Recognition (FER) can roughly be divided into appearance and geometric-based. The appearance features are extracted from facial image intensities to represent a discriminative textural pattern while the geometric ones need accurate landmark positions from which different relations can be constructed. The geometric features are, however, very sensitive to the individual face shape configuration and are therefore less consistent in person independent scenarios.

Well known and widely successful hand-crafted features such as variations of Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), Gabor filters and Local Phase Quantization (LPQ) descriptors have also been considered for FER. While most approaches considered a regular grid of patches or the whole face region for feature extraction, there have been advances in determining common and specific salient facial regions for each expression. In [9], Happy and Routray demonstrated the importance of facial landmark detection in order to find the salient patches from which they extract features.

On the other hand, a number of researchers tried to fuse different texture encoding features in order to extract complementary information that would benefit the FER. For instance, Zhang et al. used multiple kernel learning to combine two feature representations: HOG and LBP [10].

While all of the previously mentioned methods use hand-crafted and heuristically determined features, experiments with deep learning using CNN on the FER problem were recently conducted as well. However, deep learning methods have serious over-fitting problems with small datasets that are typical for FER. Lopes et al. tried different preprocessing techniques (image normalizations, synthetic samples etc.) in order to cope with the mentioned problem and were able to achieve state-of-the-art results on the CK+ benchmark dataset [11]. Even though real-time performance is claimed, a high-end GPU is needed in order to achieve it.

An additional direction of research is to integrate temporal dimension into both appearance and geometric features when working with image sequences.

5. Biometric Attributes Estimation

Biometrics refers to the problem of subject identification based on a certain unique physical characteristic (i.e. fingerprint, iris or face). On the other hand, soft biometric attributes are traits such gender, height, and eye color that provide some useful information about the subject, but are not distinctive enough to perform identification [12]. The intrusiveness of biometric systems based on fingerprint or iris recognition reduces their applicability compared to systems based on facial image analysis that do not require physical contact, subject's cooperation nor subject's attention. Three most prominent and widely researched soft biometric attributes that can be estimated from facial images are gender, age and race.

5.1. Gender classification

Gender classification is a fundamental soft biometric attribute estimation task. Due to the significance of gender attribute, availability of public face datasets with gender labels, and simplicity of the task itself (binary classification), it was a recurrent topic in early work on facial analysis.

In 1991., human performance was matched by a simple multi-layer perceptron system named SexNet [13]. By directly using pixel values as features, they achieved accuracy of 91.9% on a manually collected dataset containing 90 images.

Discriminative properties of 7 different facial regions were evaluated on a dataset containing 800 frontal faces in [14]. Periocular region was shown to be the most informative and their multi-region method based on upper face region, left eye, and nose yielded a 5% lower classification error compared to the holistic face approach.

A step further was taken in [15]. Local discriminative DNNs were applied to the most informative facial regions determined by Sobel filtering, blurring and binarization. Experiments were performed on the aligned version of LFW dataset with 13,233 images and a subset of the even more difficult Groups dataset containing 14,760 images. Evaluation on large unconstrained datasets demonstrated in-the-wild effectiveness and cross-database experiments verified the generalization capability of the proposed approach.

Despite using a simpler holistic-face approach, previously mentioned methods were outperformed by a straight-forward CNN approach trained on 500k images [16]the problem of gender recognition from face images remains difficult when dealing with unconstrained images in a cross-dataset protocol. In this work, we propose a convolutional neural network ensemble model to improve the state-of-the-art accuracy of gender recognition from face images on one of the most challenging face image datasets today, LFW (Labeled Faces in the Wild, demonstrating the power of CNNs.

5.2. Age estimation

Age estimation is one of the most challenging and broadly researched topics in the face analysis field. Early work was primarily based on geometric features, ageing pattern subspaces or manifold learning. A drawback of the mentioned approaches is that they require a well-aligned frontal faces. This section focuses on appearance-based methods that are more suited for unconstrained faces.

Age estimation can be viewed as a multi-class classification or a regression problem. Recently, the label distribution method is frequently used as it combines best of the two approaches. Age estimation typically refers to the real (chronological, biological) age estimation. Apparent age estimation is a more recent endeavor, referring to age estimation as perceived by other humans. Apparent age estimators are trained on datasets where there is no ground truth real age but instead, a group of people was guessing the subject's age.

A seminal real age estimation method based on Biologically Inspired Features (BIF) was proposed in [17]. Their variation of BIF used Gabor filters to model receptive fields and MAX and STD operations as sources of nonlinearity. PCA was used for dimensionality reduction, followed by a linear SVM in case of age classification or a support vector regressor (SVR) in case of age regression experiments. Their approach outperformed all previous work on the FG-NET benchmark, and was a basis for a number of BIF-related age estimation methods.

Comparison of hand-crafted and learned features under the same experimental settings was performed in [18]. Their exhaustive experiments showed that a simple CNN with only 2 feature-extracting convolutional layers can outperform different combinations of hand-crafted features (i.e. HOG, LBP and SURF).

Power of CNNs was further demonstrated in [19]. Their age estimation approach was based on an ensemble of 20 VGG-16 models pretrained for the task of image classification. The networks were trained for classification with 101 output neurons, each corresponding to an age from interval 0-100. The final prediction was the softmax-normalized output of those neurons, averaged over the 20 networks. An impressive error drop was achieved by additional pretraining on their large and noisy IMDB-WIKI dataset, improving the state-of-the-art by a large margin. By performing additional fine-tuning on the apparent age LAP dataset, they also achieved top score on the first edition of the ChaLearn apparent age challenge.

5.3. Race classification

In the face analysis research field, terms ethnicity and race are often used interchangeably. However, they are related to sociological and biological factors respectively. Generally, ethnicity is viewed as a cultural concept, while race refers to the person's physical appearance or characteristics and is a better suited term for classification based on facial images. Categorization to 7 commonly accepted racial groups covers more than 95\% of the world population. However, due to the scarcity of public datasets with racial annotations and good sample distribution, most of the race classification research is done on simple binary (e.g. Asian/non-Asian, White/Black) or ternary (e.g. Caucasian/African American/Asian) classification.

Following the success of deep neural networks in many other face analysis fields, Wang et al. [20] showed superior performance of their DCNN method on both binary and ternary race classification tasks. Their approach was based on CIFAR-10 CNN architecture with a n-way softmax layer. A cross-entropy loss was used during the training and the networks were trained for 3 different scenarios: (i) classification of White and Black subjects, (ii) classification of Chinese and Non-Chinese subjects, and (iii) classification of Han, Uyghur, and Non-Chinese subjects. To deal with the lack of public large-scale race analysis databases, they worked with different combinations of several public face analysis datasets and additional private datasets. For all 3 scenarios, they reported classification accuracies from 99.4% to 100%.

6. Conclusion

Recent progress in development of Face Analysis (FA) technologies created an opportunity for many new innovative commercial application fields. Like in many other computer vision fields, trend towards adoption of CNNs and deep learning is obvious, but in many cases inference speed and memory requirements are neglected. Additionally, the lack of dedicated large-scale datasets becomes more obvious due to the overfitting problems. In most cases, FA methods focus on estimation of a single attribute from a single image. Integrating the temporal dimension and solving multiple tasks jointly could increase the algorithms performance.

References

[1] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[2] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," *Lect. Notes Comput. Sci.*, vol. 8692 LNCS, no. PART 4, pp. 720–735, 2014.

[3] W. Liu *et al.*, "SSD: Single shot multibox detector," *Lect. Notes Comput. Sci.*, vol. 9905 LNCS, pp. 21–37, 2016.

[4] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," *Microsoft Res.*, no. June, p. 17, 2010.

[5] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.

[6] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 177–190, 2014.

[7] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 1078–1085, 2010.

[8] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion.," *J. Pers. Soc. Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[9] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015.

[10] X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using l p -norm MKL multiclass-SVM," *Mach. Vis. Appl.*, vol. 26, no. 4, pp. 467–483, May 2015.

[11] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.

[12] C. B. Ng, Y. H. Tay, and B. M. Goi, "A review of facial gender recognition," *Pattern Anal. Appl.*, vol. 18, no. 4, pp. 739–755, 2015.

[13] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski, “Sexnet: A neural network identifies sex from human faces,” *Adv. Neural Inf. Process. Syst. 3*, no. July, pp. 572–7, 1991.

[14] L. L. L. Lu, Z. X. Z. Xu, and P. S. P. Shi, “Gender Classification of Facial Images Based on Multiple Facial Regions,” *WRI World Congr. Comput. Sci. Inf. Eng.*, vol. 6, pp. 48–52, 2009.

[15] J. Mansanet, A. Albiol, and R. Paredes, “Local Deep Neural Networks for gender recognition,” *Pattern Recognit. Lett.*, vol. 70, pp. 80–86, 2016.

[16] G. Antipov, S. A. Berrani, and J. L. Dugelay, “Minimalistic CNN-based ensemble model for gender prediction from face images,” *Pattern Recognit. Lett.*, vol. 70, pp. 59–65, 2016.

[17] G. Guo, G. Mu, Y. Fu, and T. S. Huang, “Human age estimation using bio-inspired features,” *2009 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work. CVPR Work. 2009*, pp. 112–119, 2009.

[18] I. Huerta, C. Fernández, C. Segura, J. Hernando, and A. Prati, “A deep analysis on age estimation,” *Pattern Recognit. Lett.*, vol. 68, pp. 239–249, 2015.

[19] R. Rothe, R. Timofte, and L. van Gool, “Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks,” *Int. J. Comput. Vis.*, pp. 1–14, 2016.

[20] W. Wang, F. He, and Q. Zhao, “Biometric Recognition,” vol. 9967, pp. 176–185, 2016.

Ivan Krešo, Petra Bevandić, Marin Oršić, Siniša Šegvić

Convolutional Models for Segmentation and Localization

University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia

Abstract

The revival of deep models has profoundly improved the accuracy of image classification models and provided a large improvement potential in related computer vision tasks. Recently, much attention has been directed towards dense prediction models which produce distinct output in each image pixel. This paper addresses two particular instances of dense prediction: object localization and semantic segmentation. We briefly review the underlying operation principles, present some of our experimental results and discuss ways to analyze the success of learning and the utility of the resulting models.

1. Introduction

Recent revival of deep learning has enabled construction of multi-stage computer vision algorithms in which all stages can be trained end-to-end. Most success has been achieved with convolutional models [14] which ensure translational invariance as an essential property of vision. The resulting development has led to artificial vision systems which outperform humans in large-scale image classification [23]. This progress has been steadily followed by advances in other vision tasks. Thus, it has been noticed that semantic segmentation can be carried out by applying the same ImageNet pre-trained classification model in each pixel (cf. Fig.1). The implied computational complexity has been reduced by applying the model layerwise (as opposed to patchwise), in a convolutional manner [24].

However, it turns out that straight-forward convolutional application of a classification model results in a significant reduction of the output resolution. Consequently, a smooth transition from classification to dense prediction is hampered by strict memory limitations of contemporary GPUs as we shall show in the following sections.

2. Semantic segmentation

Semantic segmentation is a computer vision task in which we classify each image pixel into the corresponding high-level class. The ground-truth class labels are determined by the kind of the object or surface which gets projected onto the corresponding pixel. Due to being complementary to object localization, semantic segmentation represents an important step towards advanced future techniques for natural image understanding. Some attractive application fields include autonomous control, intelligent transportation systems and automated analysis of photographs and video.

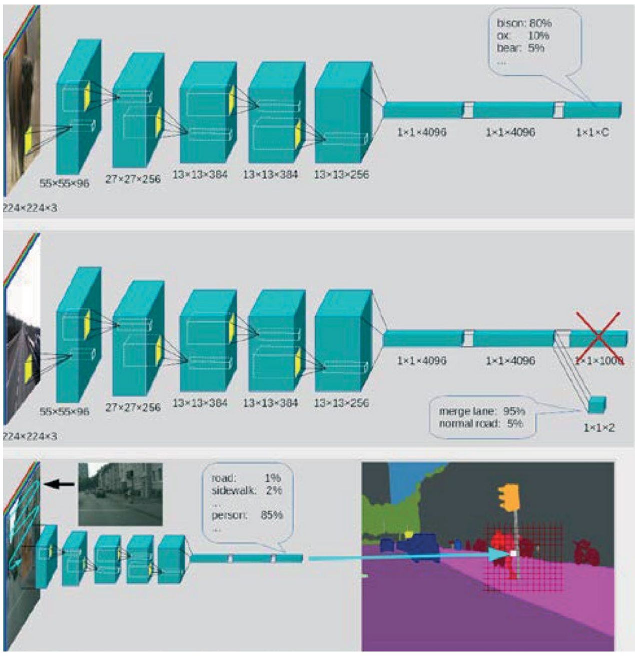


Fig. 1. A convolutional model is usually pre-trained on the ImageNet dataset which comprises 10^6 images and 10^3 classes (top). The model can be easily adapted to a simpler task by fine-tuning on the target dataset (middle). The simplest approach to achieve dense prediction would be to slide the model over all image positions (bottom). In practice we optimize this idea by applying the model *layerwise* in a fully convolutional fashion [24].

2.1. Architectural considerations

When designing an architecture for semantic segmentation we usually start with a network created for image classification. This allows to pretrain the model on the ImageNet dataset, which typically leeds to best results. In image classification task, the model output is a vector representing the distribution over classes for the whole image. In order to repurpose any classification architecture for segmentation, we need to remove the global pooling at the end and replace all fully connected layers with convolutions. However, due to intermediate pooling layers, we still get a 32x subsampled prediction. There are two ways to restore the lost resolution and get predictions at the pixel level. One way is to use intermediate feature maps before pooling layers in each block to recover lost information and refine boundaries in downsampled representation [26]. The other way is to remove pooling layers and introduce dilated convolutions [3] that will preserve the same size of the receptive field. The main downside of the dilated convolution approach is its inefficiency due to large resolution of the deep layers with many feature maps. Another downside is that we are forcing the model to propagate even very small objects through all layers of the network, which leads to potentially losing some model capacity. The upside is its simplicity because it is the easiest way to convert the network from classification to segmentation task. However, we can avoid the problems with dilated convolution and still achieve high prediction density by using ladder-style blending [26] which leverages intermediate feature maps to restore the lost details. We have successfully used this technique to successfully convert the 32x subsampled representation to the 4x subsampled output [13]. Our upsampling subnet is very efficient and introduces only a negligible increase in running time. This is achieved by blending two feature tensors from different subsampling levels with a single 3x3 convolution. It turns out that the pretrained classification network can very well adapt to this simple blending technique during fine-tuning.

2.2 Experiments on the Cityscapes dataset

We evaluated our models on the Cityscapes dataset [5] which consists of 5000 images with fine annotations and 20000 images with coarse annotations. In our experiments we used only the fine annotations. The dataset is labeled with 19 classes. The resolution of the images is 1024x2048 (cf. Fig. 2).

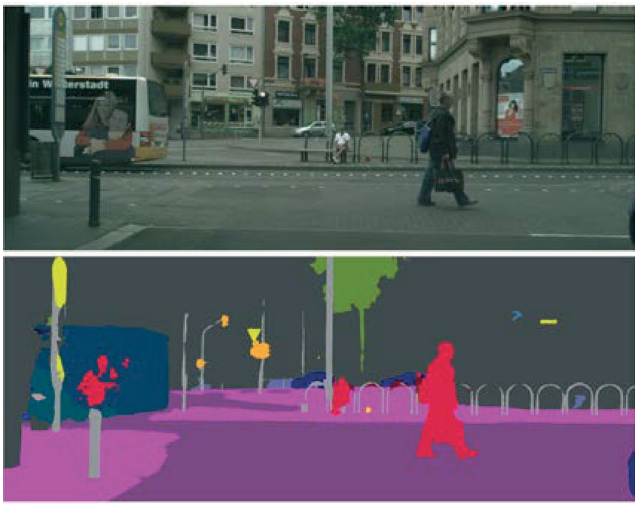


Fig. 2. Original image from Cityscapes test (top) and the dense predictions (bottom) of our semantic segmentation model (purple: road, dark blue: bus, person: red, etc). Note that a commercial sticker on the bus has been erroneously segmented as class person. Although depicting persons, that particular region should be segmented as class bus to which it semantically belongs.

The quality of semantic segmentation models is usually evaluated with intersection over union (IoU). For each class we consider pixels corresponding to the predictions and the ground-truth annotation. The IoU metric is then defined as the ratio between intersection and union of those two areas. Finally, we take the mean IoU across all classes or mIoU for short [8]. Table 1. shows our results on Cityscapes validation subset with models based on the DenseNet-121 architecture [10] pretrained on ImageNet. DenseNet 32x is the baseline model where a 32x subsampled prediction is produced right after the last DenseNet block. LadderDenseNet 4x uses ladder-style feature blending [13]. In Dilated 8x DenseNet 4x we used dilated convolutions in the last two blocks to obtain 8x subsampled prediction followed by one level of ladder-style feature upsampling to produce 4x subsampled output. Note that we couldn't use dilated convolutions to directly obtain 4x prediction due to memory limitations. LadderDenseNet 4x outperforms Dilated 8x DenseNet 4x despite requiring less memory and leading to faster execution. The large improvement between DenseNet 32x and LadderDenseNet 4x reveals the importance of prediction density. We came to similar conclusions in experiments on PASCAL VOC 2012 and CamVid datasets

3. Object localization

The purpose of object localization is to find objects of various classes in the input image and describe them with bounding boxes and class labels. This task is challenging as objects may vary in size, shape, pose, occlusion etc. Existing approaches fall into two groups. Two-stage approaches first perform class-agnostic localization of object candidates. In the second stage, the candidates are classified one at a time. On the other hand, single stage approaches produce dense predictions of bounding boxes and class labels in the compound processing step. Two stage approaches still achieve better accuracy, however we prefer single stage approaches due to simpler design and better execution speed.

3.1. Single shot detector

Single Shot Detector (SSD) [16] is the first one-stage approach to achieve accuracy comparable to two-stage approaches. It enables real time execution on 512x512 images. SSD handles the problem of varying object size by making predictions at suitable layers of a deep image representation. The features are extracted by a convolutional model consisting of the first 5 convolutional blocks from the VGG architecture [25] and 4 additional convolutional blocks. Each block subsamples the resolution of the previous block by the factor of 2. The last 6 levels of representation are connected to multibox heads which perform dense prediction of object classes and bounding box positions. Bounding box predictions are performed for multiple aspect ratios: {0.3, 0.5, 1, 2, 3}.

3.2. Experiments on MOT 2015 dataset

We evaluated the SSD approach on MOT 2015 dataset [15]. We split each training sequence into train and validation subsets such that the last 20% of images in each video are moved to the validation subset. This produces 4334 training images and 1087 validation images (we omit images that do not have any ground truth detections). The training procedure was the same as for SSD300 [16]. We experimented by adding a prediction with a taller aspect ratio (due to the fact that pedestrians are usually in a standing pose) but that did not result in any significant improvement. We display the results in Table 1. We notice a large improvement when training SSD on MOT 2015 train rather than training on 20 object classes from PASCAL VOC 2007 + 2012. Note that the competing algorithms were not tuned on MOT2015: the presented improvement is due to opportunity to better fit our model to the data. This emphasizes the importance of learning on training data whose distribution matches the distribution of the test data.

Sample detections are shown in Figure 3. SSD achieves very good accuracy on large to medium sized object while occasionally having trouble with small or distant objects. The method also has troubles with predicting false positives as well as classifying an object to a wrong but similar class (eg. mistaking a sheep for a cow). Our current experiments show that such problems can be significantly diminished with improved models. However, this research is still incomplete and so we will have to present it elsewhere.

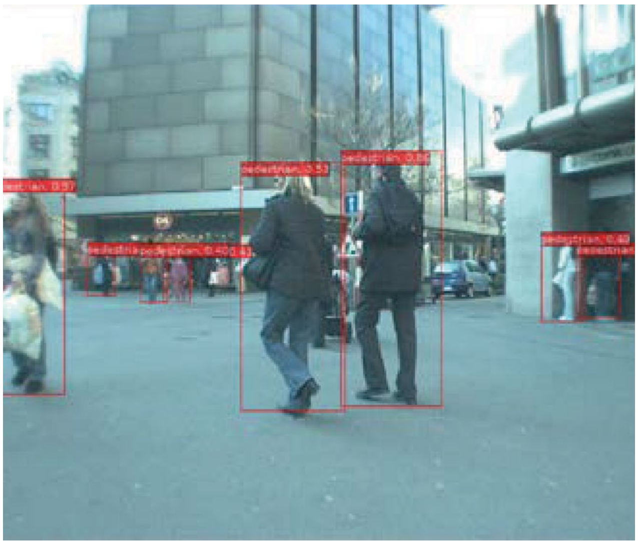


Fig. 3. Pedestrian detections on MOT 2015 val obtained by an SSD model trained on MOT 2015 train. Note that only the smallest three pedestrians have been missed.

4. Analysis of the learned models

Deep models achieve state-of-the-art performance in many computer vision tasks. However our understanding on how and why those models work remains limited. Answering these questions would not only help us improve on existing models (e.g. by understanding why deep models make mistakes), but also could play an important role in real-world application of deep models (e.g. anticipate legal implications of using deep models in practice).

4.1 Feature visualization

One way to answer how deep models work in a human friendly way is by using qualitative representations of different layers in a network. A simple example of qualitative analysis is visualization of filters in the first layer of convolutional networks. This approach is however not useful for units in deeper layers. A simple solution introduced in [erhan09icmlw] is to look for input patterns that maximize the activation of a hidden unit rather than visualizing unit content directly. Defined this way, a feature h can be visualized by locating an image patch $\mathbf{x}^*$ which maximizes its value given the model parameters ϕ:

$$\mathbf{x}^* = \text{argmax}\ \phi. \tag{1}$$

However, this definition opens up a new set of challenges [olah17distill]. For example, how to choose a hidden unit? Is it more useful to do visualize a single neuron, a single feature map, or the whole layer? Is there more than one pattern that could represent what makes a unit fire (e.g. should a neuron responsible for detecting birds fire for both penguins and hummingbirds)? Furthermore, optimizing just to make units fire does not necessarily lead to interpretable visualizations. This method can also be used to generate examples that the network classifies into one of the possible classes with a high level of confidence, without the input necessarily making visual sense to a human.

We can solve the problem of finding input patterns that maximize the activation of the hidden unit using gradient descent. We usually start from a randomly sampled image, calculate gradient of the output of the hidden unit of interest with respect to the input, and finally apply the gradient to the input. However, basic gradient descent usually gives us uninterpretable images. This problem can be solved by expanding the original problem with a suitable regularizer. Results can be further improved by slightly perturbing the input between optimizations steps to make the final visualization more robust to image transformations. The most common perturbations are blurring, jittering, scaling and rotating the input before calculating the gradient. Fig. 4 shows different types of visualization for a DenseNet architecture fine-tuned for classification on VOC 2007.

Table 1. Semantic segmentation experiments on Cityscapes val. Sx denotes S times subsampled predictions which are subsequently upsampled with bilinear interpolation. All training and evaluation images in this experiment were resized to 1024×448, while the batch size was set to 4.

Model	mIoU (%)
DenseNet 32x	62.52
Dilated 8x DenseNet 4x	71.56
LadderDenseNet 4x	72.82

Table 2. Object localization experiments on MOT 2015 val.

Model	Average Precision
SSD, ImageNet + Pascal0712 [liu16eccv]	57.06%
Agg. channel features, INRIA [dollar14pami]	60.2%
SSD, ImageNet + MOT 2015(ours)	75.5%

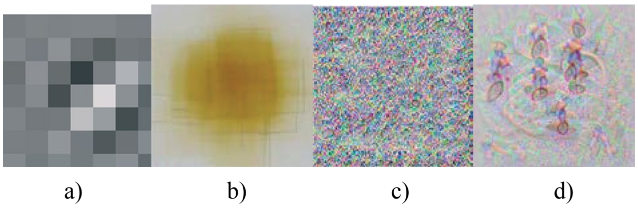


Fig. 4. Different visualizations for DenseNet 121 fine-tuned on Pascal VOC 2007: filter #20 of the first convolution layer (a), input pattern that maximizes its activation (b), input patterns that maximize the class ‘bicycle’ (c,d). No regularization was used for generating (c), while for (d) we used jittering, scaling, rotating and blurring.

4.2 Adversarial examples

Adversarial examples arise when a given image is purposively modified in a way to disrupt the correct prediction by the target model [2, 1]. If the model is not trained in a defensive manner, imperceptible perturbations can be crafted which cause the model to change its prediction away from the correct class y, while still reporting a high level of confidence. Suppose the model f provides a correct prediction in image $\mathbf{x}_i$: $f(\mathbf{x}_i) = y_i$. Then, we can recover the adversarial perturbation $\mathbf{r}$ by optimizing the following problem:

$$\min \| r \|_2 \ s.t.\ f(\mathbf{x}_i + \mathbf{r}) \neq y_i, \quad \mathbf{x}_i + \mathbf{r} \in [0,1]^m.$$

This problem can be solved by propagating the adversarial gradients to the input image $\mathbf{x}_i$ and subsequently optimizing $\mathbf{x}_i$ with gradient descent. Adversarial images can be crafted for dense prediction models as well, as shown in Figure 5.

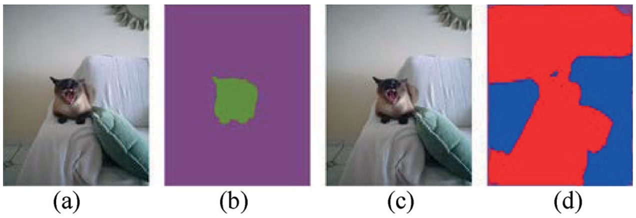


Fig. 5. Original image from Pascal VOC 2007 train (a) and the predictions (b) of our semantic segmentation model (green: cat, purple: background). Adversarial image (c) and the predictions (d) of the same model (red: dog, blue: sofa).

Following the discovery of adversarial examples, a number of exploits have been devised in literature, which, in theory, could seriously compromise practical computer vision applications. For instance, an attacker could wreak havoc in autonomous traffic by decorating stop signs with adversarial stickers [9]. However, a later study has shown that such threat could not be reproduced in more realistic localization experiments [17] where the traffic sign is observed from a variety of viewing directions. This is an important empirical finding since adversarial examples are not endemic to deep learning [22, 19]. In fact, virtually all existing vision systems based on learning (either shallow or deep) are vulnerable to adversarial attacks. Many of these systems will have to be upgraded in order to avoid successful exploits which are likely to arise in near future. Several recent papers offer interesting solutions to this problem [hinton15arxiv, cisse17icml]. Some of them are able to learn on unannotated input images which implies they could be used to support semi-supervised learning [20]. A recent defensive approach achieved almost complete resistance on CIFAR and MNIST datasets [18].

The study of adversarial examples is important even if we disregard the importance of preventing exploits. We know that existing deep models are prone to overfitting due to extremely high capacity [28]. Adversarial examples might lead us towards new regularization techniques that will improve the representation quality and further enhance the accuracy of the predictions.

5. Conclusion

We have reviewed deep convolutional models for semantic segmentation and object localization in natural scenes and presented some of our own contributions in the field. Our experiments [13] were first to confirm the utility of the recently proposed DenseNet architecture [10] for dense prediction in large images. Our model is able to restore the resolution of the dense prediction by blending higher level features at lower spatial resolution with their lower-level higher-resolution counterparts [26]. Such ladder-style blending achieves high spatial accuracy with a very lean upsampling path which significantly relaxes memory requirements and enables real-time processing of large natural images. We are currently able to process 2 Megapixel images (2048×1024) at 13 Hz on a single Titan X GPU with a model that achieves 75% mIoU on the Cityscapes test subset. Our current best result on Cityscapes test is 78.4% mIoU with a multi-resolution forward pass. These figures will improve when we complete our current experiments on the combined training dataset (fine and coarse images).

References

[1] Naveed Akhtar and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. CoRR, abs/1801.00553.

[2] Joan Bruna et al. Intriguing properties of neural networks. ICLR 2014.

[3] Liang-Chieh Chen et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40(4): 834-848 (2018).

[4] Moustapha Cissé et al. Parseval Networks: Improving Robustness to Adversarial Examples. ICML 2017: 854-863.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[6] Piotr Dollár et al. Fast Feature Pyramids for Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 36(8): 1532-1545 (2014).

[7] Dumitru Erhan et al. Visualizing Higher-Layer Features of a Deep Network. ICML Workshop on Learning Feature Hierarchie. 2009.

[8] Mark Everingham et al. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision 111(1): 98-136 (2015).

[9] Ivan Evtimov et al. Robust Physical-World Attacks on Machine Learning Models. CoRR abs/1707.08945.

[10] Gao Huang et al. Densely Connected Convolutional Networks. CVPR 2017: 2261-2269.

[11] Josip Krapac, Sinisa Segvic, Weakly-Supervised Semantic Segmentation by Redistributing Region Scores Back to the Pixels. GCPR 2016: 377-388.

[12] Ivan Kreso et al. Convolutional Scale Invariance for Semantic Segmentation. GCPR 2016: 64-75.

[13] Ivan Kreso et al. Ladder-Style DenseNets for Semantic Segmentation of Large Natural Images. ICCV Workshops 2017: 238-245.

[14] Alex Krizhevsky et al. ImageNet classification with deep convolutional neural networks. Commun. ACM 60(6): 84-90 (2017)

[15] Laura Leal-Taixé et al. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. CoRR abs/1504.01942.

[16] Wei Liu et al. SSD: Single Shot MultiBox Detector. ECCV 2016.

[17] Jiajun Lu et al. NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles. CVPR Workshop on Negative Results in COmputer Vision 2017.

[18] Aleksander Madry et al. Towards Deep Learning Models Resistant to Adversarial Attacks. CoRR abs/1706.06083.

[19] Michael McCoyd, David A. Wagner: Spoofing 2D Face Detection: Machines See People Who Aren't There. CoRR abs/1608.02128..

[20] Takeru Miyato et al. Virtual Adversarial Training: a Regularization Method for Supervised and Semi-supervised Learning. CoRR abs/1704.03976.

[21] Chris Olah, Alexander Mordvintsev and Ludwig Schubert. Feature Visualization. Distill, 2017.

[22] A. Ramanathan et al: Adversarial attacks on computer vision algorithms using natural perturbations. ICCC 2017.

[23] Olga Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3): 211-252 (2015).

[24] Evan Shelhamer et al. Fully Convolutional Networks for Semantic Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(4): 640-651 (2017).

[25] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015.

[26] Harri Valpola. From neural PCA to deep unsupervised learning. CoRR, abs/ 1411.7783, 2014.

[27] Valentina Zadrija et al. Patch-Level Spatial Layout for Classification and Weakly Supervised Localization. GCPR 2015: 492-503.

[28] Chiyuan Zhang et al. Understanding deep learning requires rethinking generalization. ICLR 2017.

Mladen Fernežir, Enes Deumić, Ivan Borko, Vjekoslav Giacometti, Dominik Šafarić, Marko Velić, Vedran Vekić, Davor Aničić

Computer Vision R&D for Classifieds in Styria Media Group

Styria Media Services ltd. Data Science Department, Oreškovićeva 6H/1, 10000 Zagreb, Croatia

Abstract

In this paper, we present two computer vision projects that were deployed as services for the Styria Media Group's classifieds: hierarchical fine-grained image categorization and image similarity search. For image categorization, we generalize the previous accuracy vs. specificity approach to automatically offer sets having the best combined accuracy and specificity, instead of returning single element suggestions. We also modify the original specificity measure to be more appropriate for the classifieds use case: minimizing the number of required clicks to reach the desired leaf category. Further, we describe our approach of utilizing a deep learning classification model for another task: creating binary descriptors in an end-to-end manner to be used for image similarity retrieval. To accomplish this task, we combine various features from different parts of the network, use multimodal learning which combines images and text from classified's ads, and finally, we employ triplet metric learning for color encoding.

1. Introduction

Styria, founded in 1869, is one of the leading media groups in Austria, Croatia, and Slovenia. As a part of the Styria Media Group, in early 2015, a team was formed to develop data science solutions for the entire group, combining natural language processing and computer vision research. Computer vision research and development for the Classifieds Project started with a clear goal to improve user experience on both the buyers' and the sellers' side of the online sales process for the Styria Group's classifieds (2nd hand marketplaces). The goal was to encourage users to do more ad placements and to have more productive searches. This would directly increase the value of the classified for its users.

For the buyers' side, the result of the project is a service called Fashion Cam, built for the Austrian Willhaben classified. The service enables buyers to find visually similar objects more easily. At first, the service was developed only for fashion but now also for furniture and antiques, with other categories soon to follow.

For the sellers' side, the end result is automatic category suggestion based on one or more images, developed for the Njuskalo classified in Croatia. The service makes the ad posting process easier and faster for the sellers.

Both products were possible due to recent advances in deep learning [1], [2], specifically in Convolutional Neural Networks (CNNs) [3]. The progress in the field was facilitated by the availability of large amounts of labeled data, modern GPU advancements, and also by hosting large-scale visual recognition competitions in the academic community based on the ImageNet dataset [4].

2. Hierarchical fine-grained image categorization

For the classifieds use case, it is common for the categories to be organized in a hierarchical manner into a specific category tree. Typically, there are multiple problems to handle: semantically similar categories in different parts of the tree, highly uneven category distributions, label quality concerns, and also, issues related to the fine-grained nature of objects to be recognized. For such fine-grained use cases, there is a problem of large intra-class variance and at the same time, small inter-class variance between some categories in the classified's categorization tree. The fine-grained problem is an active area of research tackled on diverse datasets, e.g. Oxford Flowers [6], Oxford-IIIT Pet [7], Stanford Dogs [8], CUB200-2011[9] and Cars196 [10].

At first, the problem was approached as a standard leaf classification task. The CNN network was trained to predict confidences for each of the leaf categories, using the actual leaf categories that users had chosen when placing the ads as ground truth labels for each input image. For cases where there were multiple images for the same ad, confidence predictions were averaged to obtain more accurate results.

To return the final category suggestion to our client, a separate model was trained to suggest the best subset of up to 3 nodes in the classified's categorization tree.

2.1. Architectures

The choice of the actual CNN architecture is determined by two factors: actual classification performance, and also by the required computational performance to be able to handle real-time classification requests. Currently, the models in production use elements of the GoogLeNet [11], Darknet[12] and DenseNet[13] architectures.

2.2. Revisiting the accuracy-specificity trade-off

When dealing with a hierarchical category structure, there is a possibility of returning one or more inner nodes in the categorization tree as the final category suggestion, instead of just the most confident leaf. This enables gains in accuracy, at the cost of some specificity.

Our initial solution was adopted from the Hedging Your Bets paper [14]. The paper defines a measure of specificity for each of the tree nodes, which enables joined confidence and specificity node scoring. The final category suggestion is the node having the maximum score.

Still, since there are cases of semantically similar categories in distant parts of the categorization tree, in many cases it would be best to offer all similar suggestions. The main limitation of the original HYB algorithm is that it could only offer one node suggestion. This would result in missing some of the legitimate suggestions, or moving back all the way to the common ancestor too close to the root of the categorization tree.

To solve these issues, we redesigned the original algorithm to generalize scoring to sets of nodes. This required a redefined specificity measure, which was also more appropriate for the final use case: minimizing the number of clicks that the user would have to take from our suggestions to the desired leaf category.

2.3. Categorization examples

Figures 3. and 4. showcase our category suggestions. Note that in the first case (Figure 3.) “hand tools” appear at two different places in the categorization tree. The second case illustrates a typical situation when it makes sense to offer both men’s and women’s categories (Figure 4).

3. Custom and fast visual search for real world images

For the Fashion Cam service and image similarity search in general, the biggest problem is the definition of similarity itself. There is always a semantic component, corresponding to the classified’s leaf category the ad was placed in. Other aspects are more visual: material, shape, texture, and color. In some cases, there is also the brand component which has its own important semantic and visual contributions.

The end product had to take into account both semantic and visual aspects when returning the most similar image for a given image query. At the same time, it also had to be fast to offer real-time service to our clients. Another limitation was in the available data itself which only had ad category annotations, without additional attribute tags.

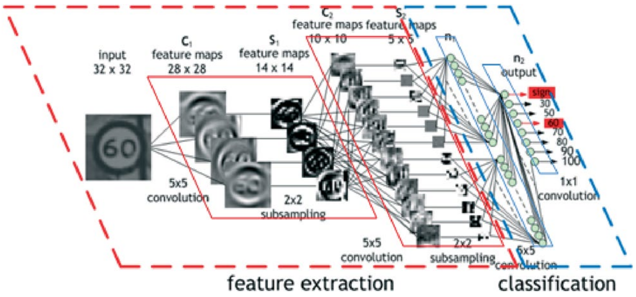


Fig. 1. Convolutional neural networks enable hierarchical learning of features: from more basic like edges and blobs to more abstract ones, enabling final object categorization.
Image by Maurice Peemen [5].

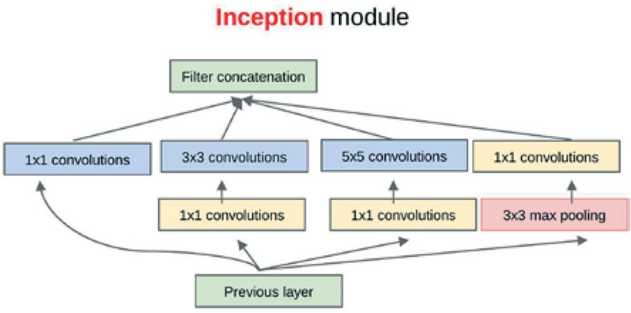


Fig. 2. Inception module, the basic component of the GoogLeNet architecture. The input layer is examined by convolutions of different kernel sizes (1 x 1, 3 x 3 and 5 x 5).

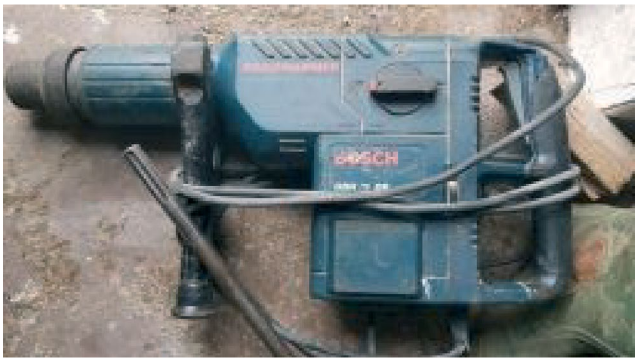


Fig. 3. Suggested tree nodes: 1. Machine and tools / Construction machinery and tools / Hand tools and tools; 2. Machine and tools / Hand tools

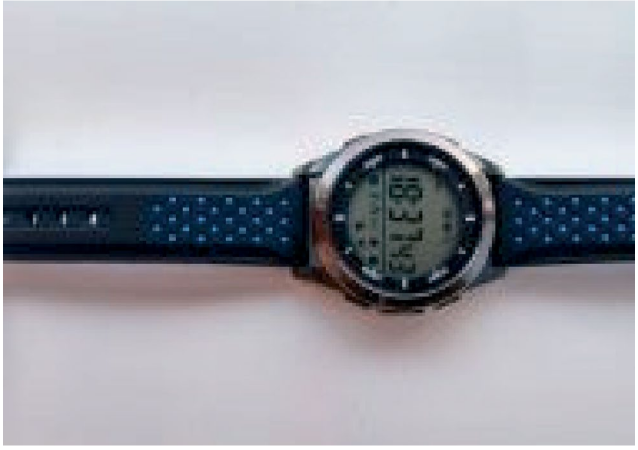


Fig. 4. Suggested tree nodes: 1. Fashion / Apparel / Watches / Men’s watch; 2. Fashion / Apparel / Watches / Smartwatch; 3. Fashion / Apparel / Watches / Women’s watch

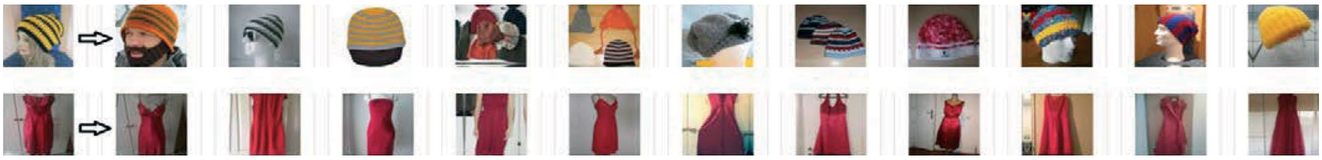


Fig. 5. Search results when using image descriptors more focused on semantics (top row) and when using descriptors with more emphasis on visual features (bottom row).

The approach we used to solve the image similarity search problem falls into the general category of representation learning [15], and more specifically, into the category of searching the appropriate hashing representation for each image with a data-dependent approach. An overview of the most recent data-dependent approaches to hashing is provided in [16]. We use a deep learning based data-dependent approach for two reasons: utilizing all specifics in the data to obtain better descriptors, and to have a fast end-to-end solution ready for real-time service for our clients.

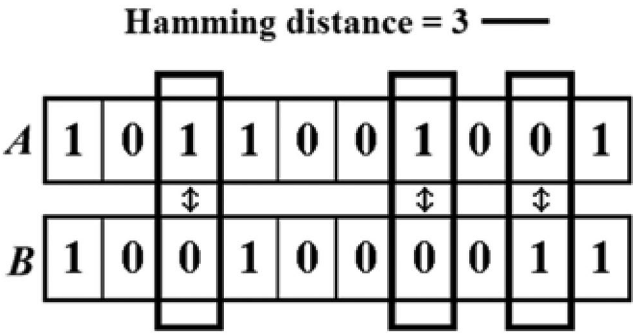
3.1 Descriptor extraction and binary encoding

The first model followed the idea presented in [17] to train a binary descriptor designed to capture category level semantics. They added an extra sigmoid fully connected layer in-between the final feature layer and the logits layer used for classification, with the idea to train that layer so that it captures high-level semantics. Two additional training loss components were used: one to make sigmoid activations close to 0 and 1, and another to make the activations as diverse as possible.

This solution was a good starting point to capture category semantics. However, it was soon discovered that we would have to do better to capture more visual aspects, especially for the fashion use case. Also, unlike [17] that used the binarized sigmoid layer for a first, coarse-level search and still reverted to a large float descriptor for fine semantic comparison, we desired a fully binarized solution to meet our run-time requirements.

To accomplish these goals of encoding both visual and semantic aspects, and to have a fully binarized descriptor, we investigated other layers in the deep neural network besides the top one meant for semantics. We took advantage of the nature of deep learning with convolutional neural networks that was mentioned in the introduction: the network learns the needed concepts hierarchically, from simpler to more abstract. The more visual aspects were present in the lower parts of the network. The final binary descriptor was formed from many different parts of the network, with many tweaks to get the satisfactory balance of semantics and visuality. Figure 5. illustrates two different combinations: a more semantically based one and a more visual one.

For faster run-time, we used a simple fully connected autoencoder to encode the final binary descriptor into a smaller one of size 64. The small one is meant for the first coarse-level search and the full one for the final ranking. All comparisons are fast on modern CPU architectures since the Hamming distance (Figure 6.) between binary descriptors can be calculated by simple XOR and bit count operations.



A	1	0	1	1	0	0	1	0	0	1
B	1	0	0	1	0	0	0	0	1	1

Fig. 6. The Hamming distance calculates the number of differing bits between two binary descriptors.

3.2 Color encoding

Color is a visual aspect that was especially important for our users. To enhance their experience, we trained a separate color encoding model and injected the color encoding layers into the main network for an end-to-end run time solution. We used triplet metric learning [18] to map perceptually similar colors in CIELAB color space to binary descriptors having similar Hamming distances.

3.3 Detecting brands

For some categories, it was especially important to be able to retrieve objects which correspond to the same brand as the query image. To accomplish this, we used a multimodal deep learning approach [19]. We used tex-

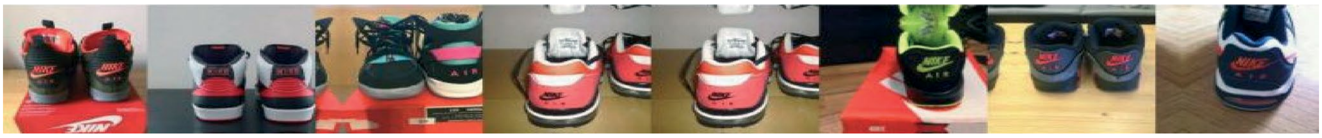


Fig. 7. Search results where brand retrieval was especially important.

tual information from the ads to detect most informative words with respect to the category in which the ad was placed. In many cases, these were brands along with some other typical words that represent types of materials. After that, the network was re-trained with this information to serve as an additional goal for learning. Results turned out to be quite good, especially for categories like sneakers or women's purses. Figure 7. illustrates similarity search results for men's Nike sneakers

4. Project results

Our response times are around 100 ms for categorization and search-by-image services, and just 50 ms when using an image that is already present on the site as a search query.

The time spent by the user in the ad insertion process, from the click on "post a new ad" until inputting text, was reduced on average by 43% from 108 seconds to 62 seconds in the current app implementation. When analyzing a subset of the data on iOS devices, where image upload and processing is much faster, the time was reduced by 71% from 89 seconds to 26 seconds. Further gains are expected after redesigning the ad placement app.

In the old process of manual categorization, the user had to do 3.1 clicks on average to reach the desired leaf category, assuming that he knew the exact path. With the new categorization service, the click path was reduced to just 0.4 clicks on average.

Customer satisfaction with the new category suggestion service was very high, with 95% of the customers rating suggestions and the whole improved user experience as excellent or very good.

The Fashion Cam project received a lot of attention from the general public and computer vision community with the biggest success of winning the best poster award at the NVIDIA GTC Conference 2017 in Munich. And most importantly, it was a well-received feature by users' feedback.

5. Conclusions and future work

Both fine-grained classification and similarity search retrieval are difficult problems to solve, even more so with data that lacks additional annotations beside the basic single-label annotations. Still, as our projects have shown, it is possible to develop both accurate and fast services to the satisfaction of the end user.

Future improvements mostly lie in the further utilization of the textual data that accompanies each ad image. For some categories, e.g. services and jobs, ad titles provide more contextual information than the images themselves. We are currently developing solutions to improve and expand the categorization service to inputs that combine both title and image, very similar to the recent advances presented in [20] and [21]. Another approach we are working on utilizes attention models for weakly supervised localization, similarly to ideas presented in [22].

To improve the similarity search service, besides the classification approaches, we are also preparing the necessary ground for similarity metric learning by using a triplet model, following [23]. Finally, we are also currently working on using user feedback to improve our similarity ranking.

References

[1] L. Deng, "Deep Learning: Methods and Applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3–4, pp. 197–387, 2014.

[2] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, vol. 187, pp. 27–48, Apr. 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, May 2017.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in 2009 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 248–255, Jun 2009.

[5] T. Dettmers, Deep Learning in a Nutshell: Core Concepts. https://devblogs.nvidia.com/deep-learning-nutshell-core-concepts/ (March 2018.)

[6] M.-E. Nilsback and A. Zisserman, "Automated Flower Classification over a Large Number of Classes," in 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp.722-729, Dec 2008.

[7] O.M. Parkhi, A. Vedaldi, A. Zisserman and C.V. Jawahar. "Cats and dogs." *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012): 3498-3505.

[8] A. Khosla, J. Nityananda, Y. Bangpeng and L. Fei-fei, "Novel Dataset for Fine-Grained Image Categorization: Stanford Dogs." (2012).

[9] C. Wah, S. Branson, P. Welinder, P. Perona and S.J. Belongie. "The Caltech-UCSD Birds-200-2011 Dataset." (2011).

[10] J. Krause, M. Stark, J. Deng and L. Fei-Fei, "3D Object Representations for Fine-Grained Categorization," 2013 IEEE Int. Conf. on Comp. Vision Workshops, Sydney, NSW, 2013, pp. 554-561.

[11] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017, pp. 6517-6525.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *2017 IEEE Conf. on Comp. Vision and Pattern Recognition* (2017): 2261-2269.

[14] L. Deng, J. Krause, A.C. Berg and L. Fei-Fei. “Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition.” *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012): 3450-3457.

[15] Y. Bengio, A. Courville and P. Vincent, “Representation Learning: A Review and New Perspectives,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798-1828, Aug. 2013.

[16] J. Wang, T. Zhang, j. song, N. Sebe and H. T. Shen, “A Survey on Learning to Hash,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 769-790, April 1 2018.

[17] H. F. Yang, K. Lin and C. S. Chen, “Supervised Learning of Semantics-Preserving Hash via Deep Convolutional Neural Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 2, pp. 437-451, Feb. 1 2018.

[18] E. Hoffer and N. Ailon, “Deep Metric Learning Using Triplet Network,” in Similarity-Based Pattern Recognition, Springer International Publishing, 2015, pp. 84–92.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A.Y. Ng. “Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011 (pp. 689-696)

[20] W. Wang, X. Yang, B. C. Ooi, D. Zhang, and Y. Zhuang, “Effective deep learning-based multi-modal retrieval,” The VLDB Journal, vol. 25, no. 1, pp. 79–101, Jul. 2015.

[21] X. He and Y. Peng. “Fine-Grained Image Classification via Combining Vision and Language.” *2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2017): 7332-7340.

[22] X.He, Y. Peng and J. Zhao. “Fast Fine-grained Image Classification via Weakly Supervised Discriminative Localization.” *CoRR* abs/1710.01168 (2017): n. pag.

[23] J. Wang et al., “Learning Fine-Grained Image Similarity with Deep Ranking,” in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1386-1393.

Dubravko Gajski[1], *Joško Jakšić*[2]

The Impact of Position and Attitude Offset Calibration on the Geocoding Accuracy of Hyperspectral Line Scanner ImSpector V9

[1]University of Zagreb, Faculty of Geodesy, Kačićeva ul. 26, 10000 Zagreb, Croatia
[2]CADCOM d.o.o., XI Trokut 5, 10020 Zagreb

Abstract

Hyperspectral images are defined as being recorded simultaneously in many, narrow, contiguous bands to provide information on the major features of the spectral reflectance of a given object. The images can be visualized as a 3-dimensional data set with two spatial and one spectral dimension and the data set is therefore often referred to as an image cube. Originally, raw hyperspectral data are combined together in an image cube with spatial, temporal and spectral dimension, after the imaging characteristic of the hyperspectral sensor (mostly push-broom scanner), and they have to be transformed to geocoded hyperspectral cube for all further spatial analysis of hyperspectral data. There are several methods to transform raw hyperspectral data (raw cube) into geocoded one. Because of imaging geometry of the hyperspectral sensor (the push-broom scanner), only the parametric geocoding methods can be applied directly. The ability of presented algorithm will be shown on test data gathered by airborne multisensor platform. The spatial accuracy of the geocoded cube will be verified on test-field.

1. Introduction

In the scientific project supported by the European Commission “Airborne minefield area reduction (ARC)”, IST–2000-25300, that lasted from year 2001. to 2003., were obtained several digital sensors and the acquisition systems and was developed the acquisition software RECORDER, [1]. Among them was purchased and later used in this project a hyperspectral line scanner V9 (ImSpector), with an insolation collecting unit (Fodis) as shown in figure 2., for the wavelengths from 430 nm to 900 nm.

The scanner was used for the acquisition of the reflectivity samples of the mine suspected areas in several different types of terrain, whereas the quality of data was limited by several factors and also was used for oil spills detection, both times as a part of the system for the Multisensor airborne reconnaissance and surveillance in the crisis and the protection of the environment. This was the reason to advance the characteristics of the airborne hyperspectral remote sensing, by use of V9, in the frame of the technological project TP-06/0007-01, in accordance with the foreseen applications [1].

There are foreseen following kinds of applications: a) measuring the *radiance* at discrete samples (static or in a direction of flight), b) measuring the *reflectance* at the discrete samples (static or in a direction of flight), c) *imaging the radiance* of the area in a form of the strip in the flight direction, d) *imaging the reflectance* of the area in a form of the strip in the flight direction. The basic measuring properties of V9 are determined by its construction. A narrow slit (8mm x 0.050 mm) at the front end of the optical system enables spectral resolution in nearly 45 channels in the wavelengths range from 430 nm to 900 nm. When the scanner is directed at nadir to the ground, the area mapped below the scanner is a narrow strip that has dimensions 0.333H x 0.00208H, where H is relative height of flight, [2]. The digital camera used for this purpose was PCO PixelFly 12bit CCD camera system with 1280x1024 pixels, pixel size 6.7 µm x 6.7 µm and scan area 8.6 x 6.9 mm, [6]. The spatial acuraccy of the imaging depend on the movements of the aerial platform, accuracy of the positioning and orientation system. While during the previous use (2001–2003) were available only GPS data, in a novel solution a positioning and orientation system is applied, combined with the parametric geocoding system program (PARGE). The advanced features of the airborne hyperspectral remote system enable wider kinds of the applications, [1].

2. Parametric geocoding

2.1. Input data for parametric geocoding

Navigation data: Position (longitude, latitude and height) and attitude (roll, pitch and true heading) stored for each line of the scanner image.

Digital elevation model: The DEM has to be given in the same coordinate system as the aircraft data.

Image/sensor general information: FOV (field of view) and IFOV (instantaneous field of view), scanning frequency, starting time, missing lines, and dimensions of the image, [3].

2.2. Geometric algorithm

The parametric processor starts with an estimate of the ‘thepretic view vector’ ($\vec{L}$) which is the imaginary line of sight to the current pixel, oriented from a horizontal aircraft facing direction north, [3]. This vector has to be set up in three dimensions to get the 'effective view vector' ($\vec{L}_t$):

$$\vec{L}_t = R \cdot P \cdot H \cdot \vec{L} \qquad (1)$$

where R, P and H are coordinate transformation matrices for the roll, pitch and true heading. The equation above describes, how the sensor is virtually turned from the north looking flight to the actual position. The vector $\vec{L}_t$ is then intersected with the DEM starting at the aircraft postion $\vec{P}_a$ to obtain the georeferenced pixel position, [3]:

$$\overrightarrow{P_{pix}} = \vec{P}_a + \vec{L}_t \frac{\Delta h}{h(\overrightarrow{L_t})} \qquad (2)$$

where Δh is the height difference between the aircraft position and the DEM intersection point. $h(\vec{L_t})$ is the height dimension of the effective view vector, [3].

2.3. Processing algorithm

- Calculate the current observation geometry; the vector ($\vec{L}$) has its origin at the entrance pupil of camera lens and at its end reaches the Digital Elevation Model (DEM)
- Find the intersection point on the surface;
- Map the image coordinates; the pixel coordinates of the image (pixel and line number) are written to an array in DEM geometry at the intersection point position.
- Gap fills; triangulation and nearest neighbor techniques are used to create a spatially continuous image

According to Schläpfer, Schaepman and Itten [3] the final processing step performs the production of geocoded images. It is separated from the main processing algorithm. This step is applied band by band which makes the processing of a band sequential raw data cube very fast.

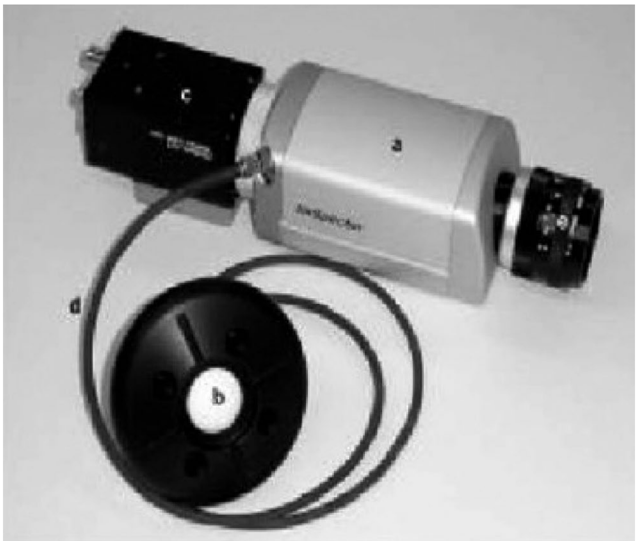


Fig. 1. ImSpector V9 with an insolation collecting unit (FODIS), [5]

3. A ground control point based offset recalibration

A ground control point (GCP) based offsets estimation tool was developed for PARGE application, [3]. The inversion of the geocoding algorithm allows the calculation of the aircraft position for each GCP. The transformed view vector is subtracted from the GCP position and stretched by the relative height:

$$\vec{P'_a} = \overrightarrow{P_{GCP}} - \vec{L_t}\frac{h_a - h_{GCP}}{h(\vec{L_t})}, \qquad (3)$$

where $\vec{P'_a}$ i $\overrightarrow{P_{GCP}}$ are the position vectors of the aircraft and the GCP, with the absolute heights h_a and h_{GCP}, [3].

The differences between estimated positions $\vec{P'_a}$ and the real navigation data are analyzed statistically to obtain the offsets. The offsets can be calculated for roll, pitch, heading, x-y navigation, height and/or field of view (FOV). The angular and distance offsets for a number of GCPs are evaluated statistically to obtain the corresponding offset estimates as follows, [3]:

- Roll: average of the angular offsets in scan direction,
- Pitch: average of the angular offsets in flight direction,
- X-Offset: average of the distance offsets in longitudinal direction,
- Y-Offset: average of the distance offsets in latitudinal direction,
- Heading: minimum correlation of the angular offsets in flight direction (pitch) to the pixel distances from nadir,
- Height: minimum correlation of the angular offsets in scanning direction (roll) to the pixel
- distances from nadir.

For heading offset estimation, the correlation between pitch offset and nadir distance is minimized by iteratively adjusting the true heading average. An analogous procedure is used for the height with the roll as indicator. Since each offset potentially depends on the others, iterations may be done between them; e.g. the heading offset may be iterated together with the pitch offset over sloped terrain, [3].

3.1. Test Field

The test field on Pula airport was used for GCP calibration procedure. The hyperspectral scanning of the test field was performed in October, 2008. Metal plates and crosses were used as signals for ground control points. The coordinates of the GCPs are determined in Gauss-Krüger metric coordinate system, 5th zone, by precise tacheometric measurements, relied on relative, static GPS-measurements. So, the accuracy of GCPs lies at the cm-level.

The handling of the auxiliary data represents the crucial issue of the whole geocoding procedure. These data consist of aircraft position and attitude. Since the absolute calibration of these data is very uncertain, GCPs are used for the offset calibration of the same data. Calibration procedure based on GCPs was performed in PARGE software.

The first step of the procedure is importing the GCPs. Typical text file with the list of GCP coordinates (image and ground) is shown in table 1, where the first two columns represent the pixel and line number (image coordinates) and the next three columns are X, Y, H (ground coordinates).

Table 1. Text file with the GCP coordinates

px	ln	X	Y	H
1092.75	616.00	5414712.40	4972979.70	147.47
719.50	617.75	5414713.13	4972966.96	147.36
720.00	628.00	5414694.49	4972965.70	147.17
877.00	628.00	5414693.90	4972971.09	147.25

Table 2. Coordinate differences, bias and variance test between the measured GCP coordinates and the ones have been read from the geocoded image before calibration

GCP	Coordinate differences		Δy-ny	vy*vy	Δx-nx	vx*vx
	Δy [m]	Δx [m]				
2	–23,70	10,20	0,58	0,34	3,35	11,22
3	–25,43	1,94	–1,14	1,31	–4,91	24,12
7	–23,80	15,81	0,48	0,24	8,96	80,26
8	–24,96	–2,81	–0,67	0,46	–9,66	93,33
11	–22,99	9,45	1,30	1,68	2,60	6,75
12	–23,87	2,67	0,41	0,17	–4,18	17,48
16	–24,44	6,38	–0,16	0,02	–0,47	0,22
17	–25,21	8,67	–0,92	0,86	1,82	3,31
20	–24,70	8,20	–0,42	0,17	1,35	1,82
22	–23,75	8,00	0,54	0,29	1,15	1,32
Σ	–242,85	68,51	0,00	5,53	0,00	239,84
	bias n_y = –24.3m	**bias n_x = +6.8m**	σ_y =	**0,78**	σ_x =	**5.16**

Table 3. Coordinate differences, bias and variance test between the measured GCP coordinates and the ones have been read from the geocoded image after calibration

GCP	Coordinate differences		Δy-ny	vy*vy	Δx-nx	vx*vx
	Δy [m]	Δx [m]				
2	–1,1	–1,1	–0,91	0,82	–1,13	1,27
3	–1,27	0,96	–1,08	1,16	0,94	0,87
7	–0,3	0,49	–0,11	0,01	0,47	0,22
8	–0,44	0,01	–0,25	0,06	–0,02	0,00
11	–0,61	–0,25	–0,42	0,17	–0,28	0,08
12	–1,13	0,93	–0,94	0,87	0,91	0,82
16	0,34	0,32	0,54	0,29	0,30	0,09
17	0,61	–0,27	0,81	0,65	–0,30	0,09
20	1,3	–0,32	1,50	2,24	–0,35	0,12
22	0,65	–0,52	0,85	0,71	–0,55	0,30
Σ	–1,95	0,25	0,00	6,98	0,00	3,84
	bias n_y = –0.2m	**bias n_x = +0.0m**	σ_y =	**0,88**	σ_x =	**0,65**

Determination of attitude offset and position offset are two crucial steps in offset calibration. These commands are performed iteratively with intention of decreasing RMS Errors (both for attitude and position). If these offsets are efficiently optimized, our next step is final geocoding procedure. After this procedure successfully finishes, we can read the coordinates of the GCPs from the geocoded image. These coordinates of the GCPs before and after calibration and those determined by tacheometric measurements are shown in table 2, where are shown coordinate differences before calibration as well as their biases and variances. In table 3 are shown the coordinate differences after calibration. Both coordinate differences, before and after calibration, are compared in table 4. It clearly shows the great improvement of spatial accuracy of the geocoded hyperspectral image after calibration.

Table 4. Comparison between the coordinate offsets before and after calibration

Coordinate differences			
before calibration		after calibration	
Δy [m]	Δx [m]	Δy [m]	Δx [m]
–23,70	10,20	–1,10	–1,10
–25,43	1,94	–1,27	0,96
–23,80	15,81	–0,3	0,49
–24,96	–2,81	–0,44	0,01
–22,99	9,45	–0,61	–0,25
–23,87	2,67	–1,13	0,93
–24,44	6,38	0,34	0,32
–25,21	8,67	0,61	–0,27
–24,70	8,20	1,30	–0,32
–23,75	8,00	0,65	–0,52

4. Conclusion

As mentioned before, the auxiliary data is the crucial component that needs to be handled in order to achieve acceptable accuracy for intended applications. For this purpose these data (aircraft position and attitude) are obtained by GPS receiver in absolute operational mode and Inertial Measuring Unit. Since the IMU achieves better accuracy over the short term and has the higher output rate than the GPS receiver, [7], this integration is used in calibration procedure for analysis of the position and attitude offsets in order to increase accuracy. The calibration procedure is based on the GCPs determined with cm-level accuracy. As we can see in table 2, there is very strong bias shown, especially on the Y-coordinates, that originates from different geodetic datum. After calibration, these strong biases on both axes are taken into account and their impact on the geocoded image is eliminated. Thus, a great improvement in accuracy after calibration is achieved, which now approximately lies at the m-level. Better accuracy can be reached by using of the more accurate GPS-receiver and applying the Kalman-filter on the IMU/GPS integration, [7].

Acknowledges

This research was supported by technological project funded by Croatian Ministry of Science (TP-06/0007-01)

References

[1] S. Semanjski, D. Gajski, M. Bajić, *Transformation of the Hyperspectral Line Scanner into a Strip Imaging System*

[2] ZEUTEC OPTO-ELEKTRONIK GmbH, 2003, *ImSpector user manual Ver.2.2*

[3] D. Schläpfer, M. E. Schaepman, K.I. Itten, 1998, *PARGE: Parametric Geocoding Based on GCP-Calibrated Auxiliary Data*

[4] ReSe, RSL, 2006, *PARGE User Manual, Version 2.3*

[5] D. Gajski, *Pozicijsko orijentacijski sustav, parametarsko geokodiranje (RC 410)*

[6] The COOKE Corporation, 2002, *PixelFly – Operating Instructions*

[7] S. Kocaman, 2003, *GPS and INS Integration with Kalman Filtering for Direct Georeferencing of Airborne Imagery*

Slavenka Petrak, Maja Mahnić Naglić, Dubravko Rogale

Computer Technology in Fashion Design and Product Development

University of Zagreb Faculty of Textile Technology, Prilaz baruna Filipovica 28a, 10000 Zagreb, Croatia

Abstract

The computer technologies have a crucial role not only in the processes of textile and fashion design and the related competitive business, but also in fashion design education. In the following text we briefly discuss several key steps during the fashion design. For each step discussed, the indispensable usage of various systems and methods is pointed out.

1. Introduction

Over the decades, computers and fashion have developed gradually, changed with time, taste and trend. The role of innovative computer technologies in the processes of textile and fashion design is one of the indispensable factors in successful and competitive business of textile and clothing manufacturers. The dynamics of changes in fashion trends and an increasing interest in the clothing market that will reflect a person's fashion identity creates a need for designers to express their creative potential in accordance with customer needs. The application of CAD systems and software packages intended for textile and fashion design with the three-dimensional visualization of a model significantly accelerates the development of new fashion collections, whereby the realistic presentation of a designer's idea is achieved. The analysis of design fit for the selected body type is thus made possible. Also anthropometrical measurements as basis for clothing construction became more precise, faster and efficient with use of modern 3D body scanning technology, encouraging individual approach in designing and creating unique garments or small made-to-measure collections.

2. Role of ICT in fashion design education

Although, most designers and fashion design colleges initially use traditional design methods, including hand drawing and manual flat pattern construction, cutting-edge education focuses on computer-aided methods of design. Computer-aided design (CAD) is the use of computer technology for the process of design and design-documentation. CAD may be used to design curves and figures in two-dimensional (2D) space; or curves, surfaces, and solids as three-dimensional (3D) objects. CAD allows designers to view designs of clothing on virtual models in different sizes and with application of various colors and textures, thus saving time by requiring fewer adjustments of prototypes and samples later. It is also possible to design textile products for other fields, like automotive and furniture industry, Fig. 1. Introducing this technological aspect can help students to understand designing process a lot better and to release their creativity to the maximum [1].

3. 3D Body Scanning

The application of the 3D body scanner has an increasing implementation in the field of body measurement for garment construction [2]. Beside the linear body measurements that are most commonly used data in the clothing industry, 3D scanning is used to obtain data on body shape, anthropometric relationships of individual body parts, deviations from the normal proportions and body posture characteristics [3,4]. In this manner, all relevant data necessary for Computer-aided design and modification of garment patterns according to the individual body anthropometric characteristics are determined. International standard ISO 20 685 has been developed to ensure the comparability of body measurements defined by ISO 7250 (Basic Human Body Measurements for Technological Design) and ISO 8559 (Garment Construction and Anthropometric Surveys-Body Dimensions) obtained using various 3D body scanners.

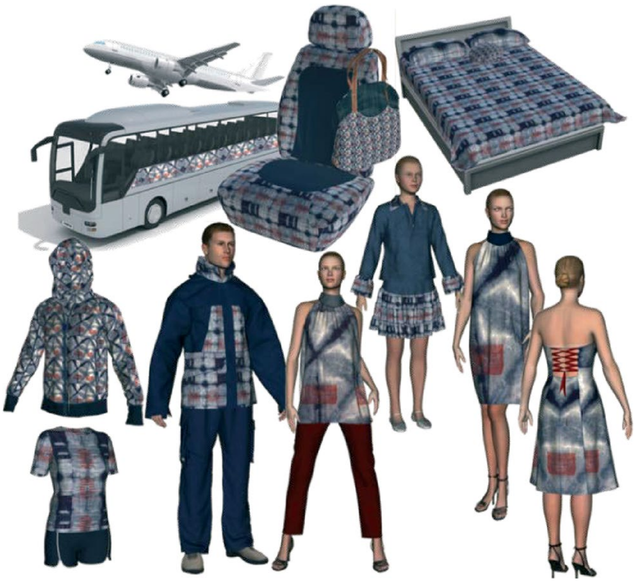


Fig. 1. 3D prototypes of clothing, in the furniture and automotive sector, in various colours and patterns

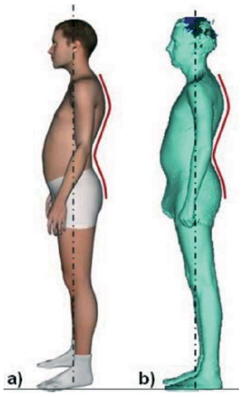


Fig. 2. Body posture adjustment: a) avatar, b) scanned body model

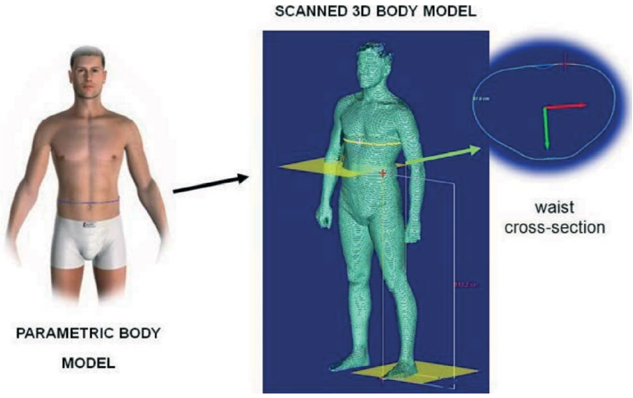


Fig. 3. Harmonization the waist girth and determination of the cross-section at the given position

However, for performing 3D garment simulations it is also necessary to ensure harmonization between the anthropometric measurements determined by 3D body scanner and the corresponding measures of an avatar in the CAD system, Fig. 2. and Fig. 3. Most existing computer programs for 3D virtual garment use parametric human models or avatars, with different number of body measurements that can be interactively customized [5].

The Vitus Smart 3D body scanner installed at University of Zagreb Faculty of Textile Technology allows users to scan an object in the area of 1,200 x 800 mm and 2100 mm in height. Scanning is performed by the system of 8 cameras and lasts 10 seconds, whereby 500,000 to 600,000 spatial coordinates of the scanned body are extracted. Data processing takes about 40 seconds [6]. Softwares ScanWorx or Anthroscan are used for human body measurements, necessary for the implementation in the computer program for the garment pattern alteration according to the individual characteristics. The use of a 3D scanners and accompanying computer program also enables precise 3D body model measurements in dynamic postures, Fig. 4., where the dimensions of the surface parts and segments volumes can be determined in order to achieve high garment fit and to ensure the comfort in dynamic conditions of use.

4. 3D flattening method for designing tight fit clothing

Tight-fit clothing items represent a specific group of products intended for wearing close to the body. When using 3D flattening method for 3D construction of tight-fit clothing it is necessary to take in consideration physical and mechanical properties of the material from which the clothing will be made. It requires comprehensive knowledge of fabric behaviour, tensile and shear properties, its behaviour on the body as well

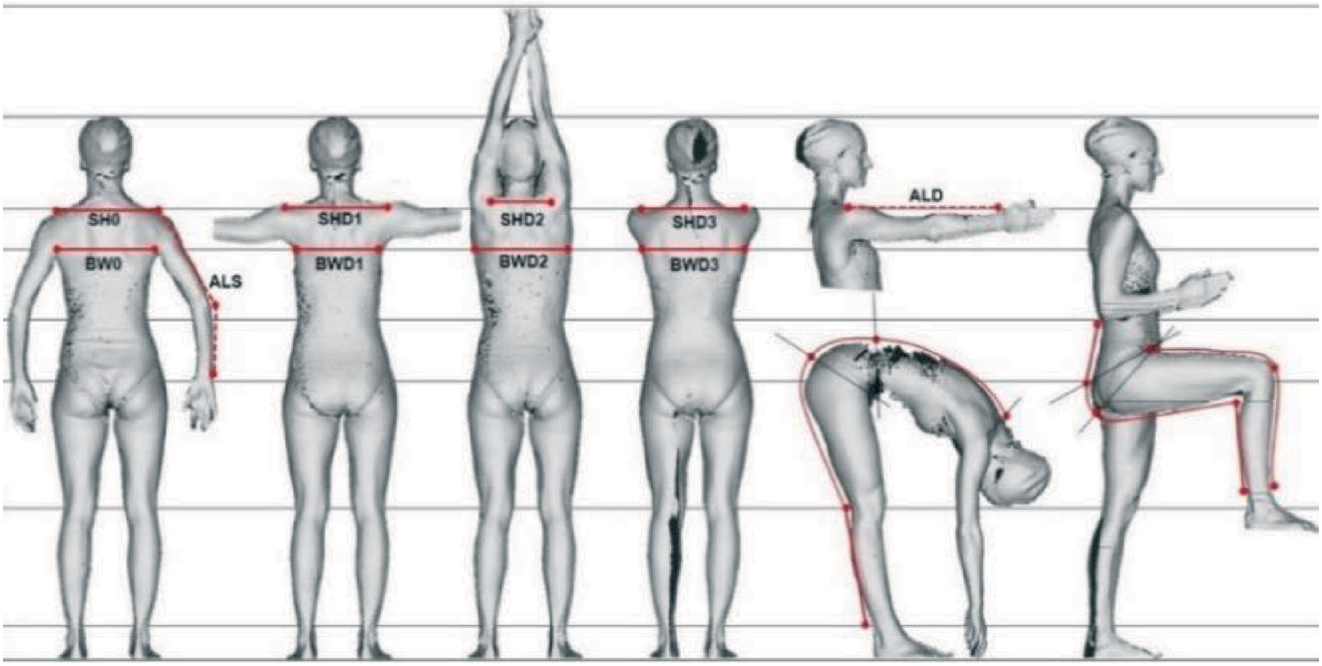


Fig. 4. Position of taken measurements in static and dynamic position

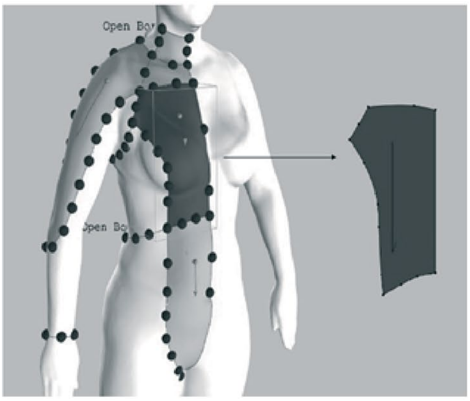


Fig. 5. Computer-based 3D clothing construction

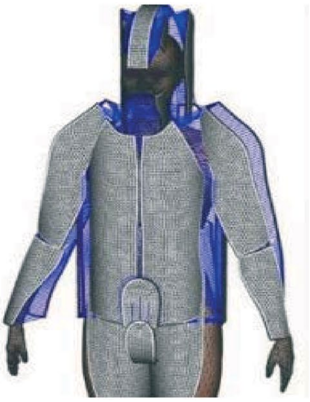


Fig. 6. 3D simulation process

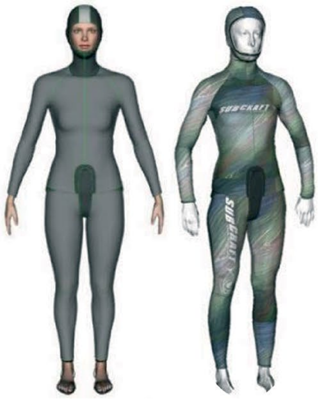


Fig. 7. Computer prototype

as constructional and functional requirements imposed on the clothing. Application of 3D flattening method for the construction of a female diving suit [7], involves drawing and creating pattern lines directly on the surface of a computer body model, separation of discrete 3D surfaces and transformation into 2D cutting parts, Fig. 5.

5. 3D simulation of model prototypes

Virtual garment simulation is the result of a large combination of techniques that have dramatically evolved during the last two decades [8]. Besides the mechanical models used within existing mechanical engineering for simulating deformable structures, many new challenges arise from versatile nature of textile fabrics. Therefore, garment simulation is based on the development of the efficient mechanical simulation models, which support the reproduction of the specific non-linear mechanical properties of textile materials. In addition, the garments interact strongly with the body, as well as with other garments layers. This requires the development of the advanced methods efficiently detecting the geometrical contacts constraining the behaviour of the fabric and integrated them into the mechanical model [9,10]. In order to verify the patterns developed by flattening method, 3D simulations of diving suit models have been performed with physical and mechanical properties of neoprene material applied to the patterns, Fig. 6 and 7.

Analysis of computer prototype showed positive assessment of 3D flattening method application for obtaining precise garment pattern suitable for production of a real prototype [11]. Material stretch analysis on computer prototype and verification of real garment prototype by professional female diver in conditions of use, confirmed that method enables development of functional tight-fit clothing.

References

[1] Petrak S., Mahnic M., Rogale D. & Ujevic D. (2011): "Computer Design of Textile and Clothing Collection – Assumption of Contemporary Remote Business", Proceedings of 11th World Textile Conference AUTEX 2011, Mulhouse, France, 1162-1168.

[2] D'Apuzzo N., (2009): "Recent advances in 3D full body scanning with applications to fashion and apparel", Proceedings of 9th Conf. on Optical 3D Measurement Techniques, Vienna, Austria

[3] Fischer T. et al. (2016): "Automatic morphological classification with Case-Based Reasoning", Proceedings of the 3rd International Conference on 3D Body Scanning Technologies, Lugano, Switzerland, 148-158, doi:10.15221/16.148.

[4] Mahnic Naglic M. & Petrak S. (2017): "A method for body posture classification of three-dimensional body models in the sagittal plane", Textile Research Journal, Online first, 1-17.

[5] Petrak S., Mahnic M. & Ujevic D. (2012): "Research of 3D Body Models Adjustment Based on Anthropometric Data Determined by Laser 3D Scanner", Proc. of the 3rd Int. Conf. on 3D Body Scanning Technologies, Lugano, Switzerland, 115-126.

[6] URL: VITUS - 3D body scanner, URL: http://www.vitus.de/, *Accessed*: [12. 02. 2018.].

[7] Zhang Y. et al. (2016): "Optimal fitting of strain-controlled flattenable mesh surfaces", The Int. Journal of Advanced Manufacturing Technology, 87(9-12), 2873-2887, ISSN 0268-3768.

[8] Magnenat-Thalmann N. (2010): "Modeling and Simulating Bodies and Garments", Springer, ISBN 978-1-84996-262-9, London, England.

[9] Fan J., Yu W., Hunter L. (2004): "Clothing appearance and fit: Science and Technology", Woodhead Publishing Limited and The Textile Institute, ISBN 1 85573 745 0, Cambridge, England.

[10] Naebe M. et. al. (2013): "Assessment of performance properties of wetsuits", The Journal of Sports Engineering and Technology, 227(4), 255-264, ISSN 1754-3371.

[11] Mahnic Naglic, M., Petrak, S., Gersak, J. and Rolich, T. (2017): "Analysis of dynamics and fit of diving suits", IOP Conference Series: Materials Science and Engineering 254. doi:10.1088/1757-899X/254/15/152007.

Matea Đonlić[1], Tomislav Petković[1], Tomislav Pribanić[1], Stanislav Peharec[2]

Structured Light 3D Body Scanner for Back Surface Analysis

[1]University of Zagreb, Faculty of Electrical Engineering and Computing, 10000 Zagreb, Croatia
[2]Peharec Polyclinic for Physical Medicine and Rehabilitation, 52100 Pula, Croatia

Abstract

In this paper we present our 3D scanning system for the surface reconstruction of the entire human body. Scanner is based on the structured light fringe profilometry approach and is assembled from only few basic components – three projectors and six cameras. These components are grouped in three scanning units which can simultaneously illuminate person from all sides and create a full 3D body model. On the application side, we propose the analysis of the human torso (back surface) with a goal to assist the detection and assessment of possible spinal deformities or other muscular changes present on the back surface. We proposed a method for the automatic detection of the symmetry curve based on the analysis of local deviations of surface curvatures. Presented results demonstrate that proposed method is comparable with commercial 3D systems.

1. Introduction

Structured light (SL) profilometry is one popular approach to the 3D surface reconstruction. The object's surface is illuminated using a pattern with a coded structure and then captured by one or more cameras. The decoding between projected and captured pattern structure enables the reconstruction of the scanned object. In order to reconstruct the entire surface of the human body, different body-regions can be either illuminated sequentially (by employing projectors in turns) [1][2] or simultaneously [3]–[5] where multiple projectors illuminate the subject at the same time from different positions. Downside of the sequential multi-projector SL reconstruction is extended scanning time (that grows linearly with the number of projectors used) meaning that subject must remain still for longer time which may be uncomfortable. On the other hand, simultaneous multi-projector SL reconstruction approaches must efficiently solve the problem of the inter-projector interference, which usually requires projecting specifically designed patterns and consequently using some nonconventional decoding methods [3][4]. Fringe projection profilometry is very popular SL approach for 3D surface reconstruction, because it's robustness – it is insensitive to ambient illumination, limitations on the colour of the scanned object are practically negligible, it can produce high-resolution scans and, depending on the decoding method, the relative positioning of the camera and the projector is not strictly conditioned (having in mind that they must retain a common FOV). In order to preserve the robustness property of the aforementioned fringe projection profilometry, in our opinion the method of choice for the surface reconstruction is temporal multiplexing approach proposed in our previous work [5]. This approach does not impose a limit on the number of projectors used nor on theirs placement and thus enables construction of complex scanners with no blind spots. The only prerequisite for previously mentioned simultaneous projection approach is the synchronization of projectors and cameras.

The second part of the paper describes our analysis of the reconstructed human body. Proposed analysis is focused on the human torso, i.e. the analysis of the human back surface topography and geometry in order to detect and assess the possible deformities of the human spine or other muscular bulges that change the symmetry in patient's posture and consequently the surface of the back. Bearing in mind that, in healthy subjects, the spine is one of the main indicators of the back surface symmetry, the first step in most of analysis methods is the estimation of the spinal curve on the back surface. The easiest approach is using reflective adhesive markers and marking a certain number of the vertebrae and then interpolate the spine curve, but more recent methods propose using a (semi-) automatic detection of the asymmetry curve. There are many different approaches to this problem – analysing the depth of the surface profiles, finding maxima in the computed surface curvature [6], or defining an asymmetry function as a left-right differences of the surface curvature [7] or surface normals [8] distribution over the horizontal profiles of the human back. Our work extends the idea of [7], using multi-scaling of the asymmetry function which effectively filters minor asymmetries inconsistent over multiple scales.

2. Multi-Projector 3D Scanning System

Our 3D scanning system is based on multi-projector multi-camera temporal multiplexing fringe projection profilometry. Fringes used for each projector are carefully designed – temporal phase shifts of each fringe set are selected to form an orthogonal basis of the discrete Fourier transform (DFT). That means that for each of P projectors we generate a set of $N \geq 2P+1$ pattern images:

$$I_{\mathrm{PRJ},k}\left(x_{\mathrm{PRJ}}, y_{\mathrm{PRJ}}\right) = \frac{1}{2} I_0 \left(1 + \cos\left(\omega_k x_{\mathrm{PRJ},k} + \varphi_k\left[n\right]\right)\right), \quad (1)$$

$\varphi_k[n] = 2\pi kn/N$, $n = 0, \dots , N - 1$ and $k = 0, \dots , P - 1$.

Using a multi-projector configuration, each camera captures N frames where each frame includes ambient illumination along with the contribution of each projector:

$$I_{\mathrm{CAM}}\left(x_{\mathrm{CAM}}, y_{\mathrm{CAM}}\right) = I_{AMB} + \frac{1}{2}\sum_{k=0}^{P-1} h_k I_k \left(1 + \cos\left(\omega_k x_{\mathrm{PRJ},k} + \varphi_k\left[n\right]\right)\right) . (2)$$

For each camera we decode a set of N images in a following manner. Firstly, we decompose each set using the Fast Fourier transform as explained in [5]. By comparing the magnitude of the k-th spectral component to some preselected threshold, we determine the area illuminated by the k-th projector and afterwards the wrapped phase ϕ_k of k-th projector can be retrieved as the negative phase of the k-th spectral component. Here, we omit details regarding wrapped phase definition but it can be found in our previous paper [5]. Unwrapping of the wrapped phase can be done using any of the unwrapping algorithms [9], and the final 3D reconstruction is obtained with the triangulation using corresponding camera projector coordinate pairs. Additionally, our post-processing includes some filtering and creating a mesh for the better visual representation of results.

An important part of this reconstruction approach is the adequate synchronization between cameras and projectors. Based on our previous work [10], we chose the software synchronization (opposed to the more expensive hardware synchronization) which relies on the precise timing of projection and acquisition steps.

3. Back Surface Analysis

Although our 3D scanning system can produce a 3D reconstruction of the entire human body we focused our 3D analysis only on the back surface. As stated in the introduction, an important part of the assessment of possible spinal deformities is the estimation of the spinal curve or some other correlated curve on the surface of the back.

Our proposed method is based on the analysis of distributions of surface curvatures and on redefining an asymmetry function. The input to the procedure is a point cloud with associated surface normals representing the back side of the subject's torso. This can be achieved using some 3D body segmentation method or manually by selecting the region of interest. Using precomputed surface normals and the reconstructed dense point cloud, surface curvatures (principal curvatures and principal directions) can be estimated [11].

Hierholzer [7] defined the surface asymmetry function as a sum of local deviations of the surface curvature in some predefined neighbourhood of the chosen point. We adopted this definition, and for each horizontal slice of the 3D back surface, we computed the symmetry function in every point interpolated over that slice. The result is two-dimensional symmetry function map. A valid theoretical assumption is that the symmetry function will achieve maximal values at points which represent the *symmetry curve*. However, the symmetry map produces many local maxima which need to be filtered in order to achieve automatic detection of the symmetry curve. Therefore, we propose *multi-scaling* of the symmetry function – accumulation of multiple symmetry functions computed over different neighbourhoods, which effectively filter minor symmetries which are inconsistent over multiple scales.

4. Results and Discussion

Our 3D scanning system is comprised of three units – each equipped with one projector and two cameras. One projector is Canon LV-WX310ST and two are Acer S1383WHne. All cameras are PointGreys' Grasshopper3 GS3-U3-23S6C. Four cameras are equipped with Fujinon HF12.5SA-1 lenses and two are equipped with Kowa LM8JCM lenses. With the maximal speed of 20 FPS one recording takes about 0.7s, but for more robust and higher quality reconstructions we propose using three frequencies (ω) with seven shifts (φ) per frequency (N = 42) for each projector coordinate which results in acquisition time of 2.1s. The 3D scanning system setup is shown in Figure 1. We used a double-sided calibration board with circular hexagonal grid pattern

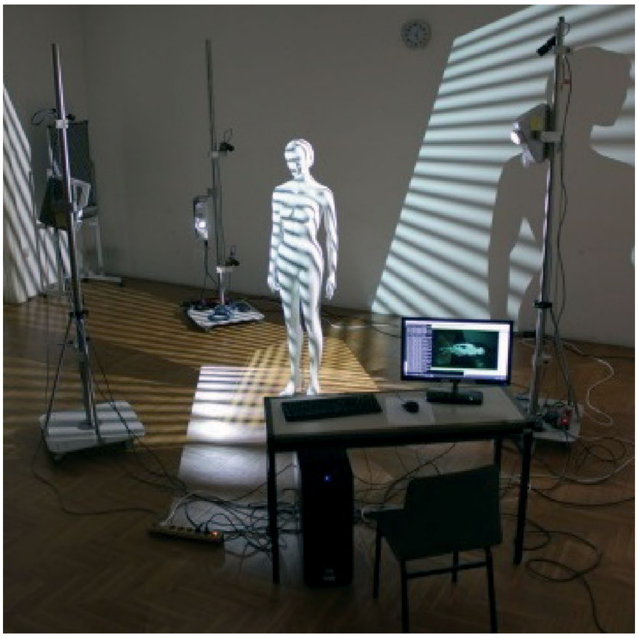


Fig. 1. Our 3D scanning system. Note the inter-projector interference pattern on the mannequin and on the floor

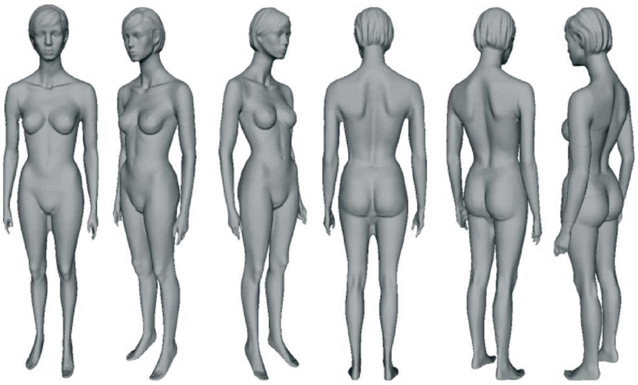


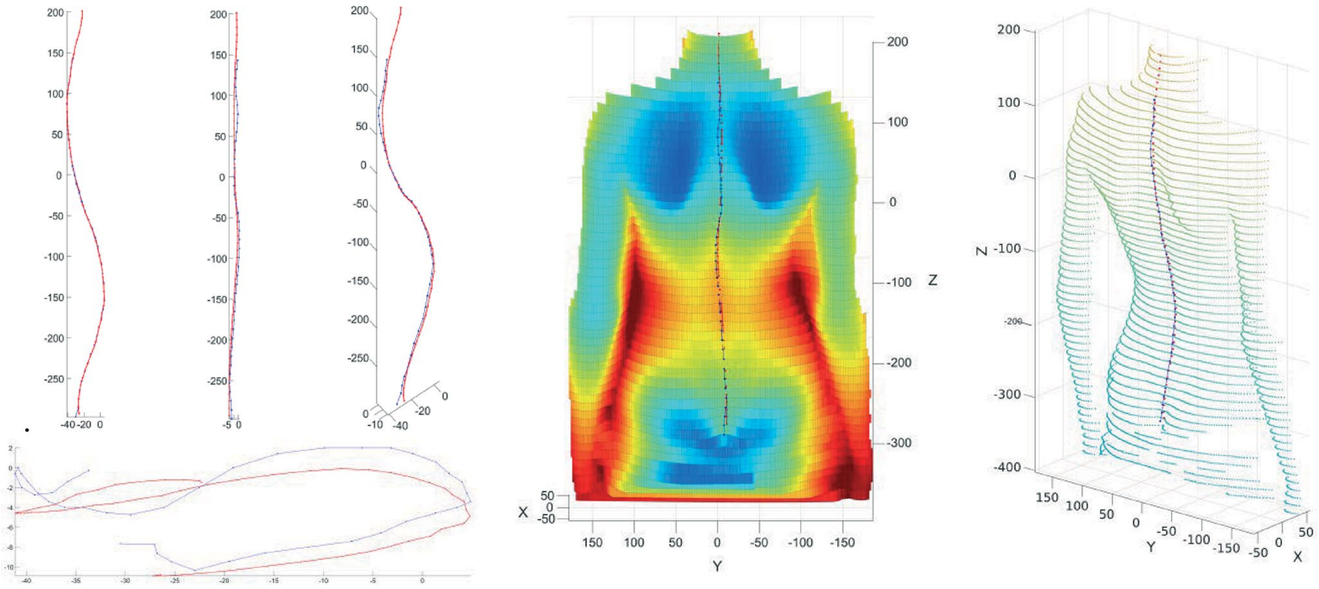
Fig. 2. Resulting 3D mesh of a mannequin after final post-processing.

and coded markings in order to achieve simple and fast geometric calibration of all three scanning units. An example of the final surface reconstruction for the scanned mannequin is shown in Figure 2. The 3D mannequin model is pictured as a mesh surface for better visual representation.

We compared the proposed method for the back surface analysis with a commercial system for the 3D spine and posture analysis – Diers Formetric [12]. We used point clouds reconstructed using a Diers system as our input and applied the proposed method for the detection of the symmetry curve. The comparison with the output of the Diers system (so-called *cls* curve) showed that methods are comparable within limits of physicians' palpation error (5mm) as shown in Figure 3. The contribution of our proposed multi-scaling of symmetry functions is presented in Figure 4. Using this approach we are able to compute the symmetry curve without using any predefined models for the curvature of the human spine.

5. Conclusion

We have proposed a 3D human body scanning system which can be used for the analysis of the human torso, specifically for the back surface analysis. The experi-



(a) (b) (c)

Fig. 3. Comparison of the proposed method (blue line, *symmetry curve*) and Diers Formetric resuts (red line, *csl curve*) in the back shape analysis. (a) Comparison of detected curves in different views (axial, sagittal, coronal, and from side). (b) Detected lines plotted on the depth map of the back surface. (c) Detected lines plotted over the input point cloud obtained with Diers Formetric system.

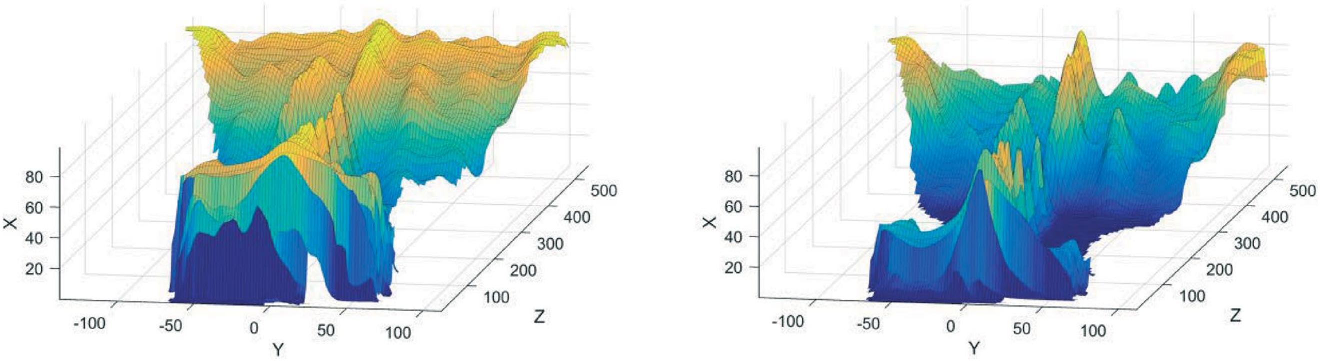


Fig. 4. Symmetry map generated using one scale (left) and using the proposed *multi-scaling* approach (right). Note how local maxima representing the "true" symmetry curve are much more prominent.

ments show that our scanner is very robust and can collect data for a dense 3D reconstruction of the entire human body in only two seconds. The reconstructed surface of the human back can then be analysed using proposed method for the detection of the symmetry curve which does not require any predefined models thanks to the proposed multi-scaling approach. The results are very promising for further extensions of the method because current results are already comparable with the commercial 3D systems.

References

[1] R. R. Garcia and A. Zakhor, “Markerless Motion Capture with Multi-view Structured Light,” *Electron. Imaging*, vol. 2016, no. 21, pp. 1–7, Feb. 2016.

[2] W.-H. Su et al., “Projected fringe profilometry with multiple measurements to form an entire shape,” *Opt. Express*, vol. 16, no. 6, p. 4069, Mar. 2008.

[3] S. Woolford and I. S. Burnett, “Toward a one shot multi-projector profilometry system for full field of view object measurement,” in *2014 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 569–573.

[4] R. Furukawa et al., “Multiview projectors/cameras system for 3D reconstruction of dynamic scenes,” in *2011 IEEE ICCV Workshops*, 2011, pp. 1602–1609.

[5] T. Petkovic, et al., “Efficient Separation Between Projected Patterns for Multiple Projector 3D People Scanning,” in *2017 IEEE ICCV Workshops*, 2017, pp. 815–823.

[6] P. Poredoš, D. Čelan, J. Možina, and M. Jezeršek, “Determination of the human spine curve based on laser triangulation,” *BMC Med. Imaging*, vol. 15, no. 1, p. 2, Dec. 2015.

[7] E. Hierholzer, “Analysis Of Left-Right Asymmetry Of The Back Shape Of Scoliotic Patients,” in *Proc. SPIE*, 1986, vol. 602, no. Biostereometrics ’85, p. 266.

[8] L. Di Angelo, P. Di Stefano, and A. Spezzaneve, “A method for 3D detection of symmetry line in asymmetric postures,” *Comput. Methods Biomech. Biomed. Engin.*, vol. 16, no. 11, pp. 1–8, 2012.

[9] C. Zuo, et al., “Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review,” *Opt. Lasers Eng.*, vol. 85, pp. 84–103, Oct. 2016.

[10] T. Petkovic, et al., “Multi-Projector Multi-Camera Structured Light 3D Body Scanner,” in *Proceedings of 3DBODY.TECH 2017, Montreal QC, Canada, 11-12 Oct.*, 2017, pp. 319–326.

[11] R. C. Wilson and E. R. Hancock, “Consistent topographic surface labelling,” *Pattern Recognit.*, vol. 32, no. 7, pp. 1211–1223, Jul. 1999.

[12] “DIERS biomedical solutions.” [Online]. Available: http://www.diersmedical.com/. [Accessed: 20-Mar-2018].

Engineering Power – *Bulletin of the Croatian Academy of Engineering* Vol. 13(2) 2018 – ISSN 1331-7210

Publisher: Croatian Academy of Engineering (HATZ), 28 Kačić Street,
P.O. Box 59, HR-10001 Zagreb, Republic of Croatia

Editor-in-Chief: Prof. Vladimir Andročec, Ph.D., President of the Academy
retired Full Professor with tenure, University of Zagreb, Faculty of Civil Engineering

Editor: Prof. Zdravko Terze, Ph.D., Vice-President of the Academy
University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture

Guest-Editor: Prof. Tomislav Pribanić, Ph.D., University of Zagreb, Faculty of Electrical Engineering and Computing

Editorial Board: Prof. Vladimir Andročec, Ph.D., Prof. Zdravko Terze, Ph.D., Prof. Slavko Krajcar, Ph.D.

Editorial Board Address: Croatian Academy of Engineering (HATZ), „Engineering Power“ – Bulletin of the Croatian Academy of Engineering, Editorial Board, 28 Kačić Street, P.O. Box 59, HR-10001 Zagreb, Republic of Croatia

E-mail: hatz@hatz.hr

Graphical and Technical Editor: Vladimir Pavlić, Dipl. Eng. (GRAPA, Ltd., Zagreb) *Press:* Tiskara Zelina, Ltd., Zelina

Proof-reader: Miroslav Horvatić, MA *Circulation:* 300