COMPARISON OF TWO METHODS FOR 2D POSE ESTIMATION OF INDUSTRIAL WORKPIECES IN IMAGES – CNN VS. CLASSICAL IMAGE PROCESSING SYSTEM

C. Siegfarth, T. Voegtle, C. Fabinski

Institute of Photogrammetry and Remote Sensing (IPF), Karlsruhe Institute of Technology (KIT), Germany – Clarissa@Siegfarth.de, thomas.voegtle@kit.edu , christopher.fabinski@student.kit.edu

Commission I, ICWG I/IV

KEY WORDS: Automatic image analysis, CNN, Shape model, Industrial Application

ABSTRACT:

Today, automatic image analysis is one of the basic approaches in the field of industrial applications. One of frequent tasks is pose estimation of objects which can be solved by different methods of image analysis. For comparison two of them have been selected and investigated in this project: Convolutional Neural Networks (CNNs) and a classical method of image analysis based on contour extraction. The main point of interest was to investigate the potential and limits of CNNs to fulfil the requirements of this special task regarding accuracy, reliability and time performance. The classical approach served as comparison to a state-of-the-art solution. The workpiece for these investigations was a commonly used transistor element. As database an image archive consisting of 9000 images with different illumination and perspective conditions has been generated. One part was used for training of the CNN and the creation of a so-called shape model respectively, the rest for the investigation of the extraction quality. With CNN technique two different approaches have been realised. Even if CNNs are predestined for classification this method delivered insufficient results. In a more sophisticated approach the system learns the parameters of an affine transformation including the sought-after parameters of translation and rotation. Our experiments confirm that CNNs are able to obtain at best only a medium accuracy of rotation angles (about ±2°), in contrast to the classical approach (about ±0.5°). Concerning the determination of translations both methods deliver comparable results, about ±0.5 pixel from CNN and about ±0.4 pixel from classical approach.

1. INTRODUCTION

Today, automatic image analysis is one of the basic approaches in the field of industrial applications, among others for inspection (e.g. check of completeness, testing fluid levels of glass bottles etc.), measurements (e.g. check of defined dimensions of a workpiece, measurement of bore diameters etc.) or object detection and pose estimation (e.g. for applications in robotics and handling techniques). The latter task can be solved by different methods of image analysis. For comparison two of them have been selected and their extraction quality has been investigated in this research project: on the one hand Convolutional Neural Networks (CNN) – a relatively new technique, e.g. (Bourez, 2017; Goodfellow et al., 2016; Patterson & Gibson, 2017) – realised by the LASAGNE system (Lasagne, 2015), on the other hand a classical method of image analysis based on edge and contour extraction using the fully developed system HALCON (Halcon, 2018). The main point of interest was to investigate the potential and limits of CNNs to fulfil the requirements of this special task regarding accuracy, reliability and time performance. HALCON served as comparison to a state-of-the-art solution. As reference an image database containing 9000 images of the workpiece in different positions, rotations and illuminations had been generated.

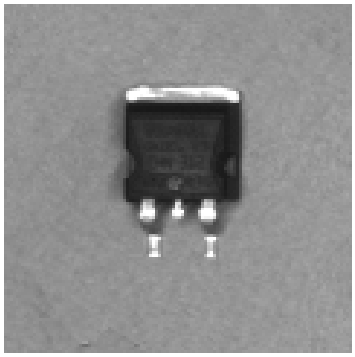

Figure 1. Workpiece (transistor element) used in this project

2. DESCRIPTION OF THE WORKPIECE

As a representative example of a workpiece for investigations and tests of CCNs and classical image analysis algorithms a transistor element was selected with a dimension of 10 mm width, 15.5 mm length and 4.5 mm thickness (Figure 1). This transistor is relatively small and therefore, challenging for detection and accurate pose estimation. Additionally, only a moderate to weak contrast to the background was chosen due to common, realistic environmental conditions in industrial production.

3. GENERATION OF THE IMAGE DATABASE

As database an image archive consisting of 9000 images has been generated. In a first step 30 basic images (pixel size = 0.25 mm in object space) had been captured orthogonal to the transistor surface with different positions of the workpiece in relation to the image area and different rotations to consider the minor perspective and illumination effects caused by the small thickness of the transistor element. Therefore, the common variability of image capture in industrial environment could be regarded. Afterwards, to generate well-defined reference images a smaller subset around the workpiece had been cut out and rotated to the initial angle ($\alpha = 0°$), so that the transistor is placed exactly in the middle of that subset (Figure 1).

In a second step each of these basic image subsets had been shifted and rotated (related to the center point of a "background" image without workpiece) to exactly (user)defined amounts by means of a classical image processing system. As rotation angles 0°, 15°, 30°, 45°, 60°, 75° and 90° had been chosen, translations in the domain of ±20 pixel. Subsequently a new subset of the same size had been cut out around the center point. In this way a lot of reference images of the same size can be created with different translations, rotations, illuminations and perspective views of the workpiece (Figure 2). For example, a basic subset had been rotated around 360° in steps of 1°. Taking additionally the translations into account an overall number of 9000 reference images had been generated in this way.

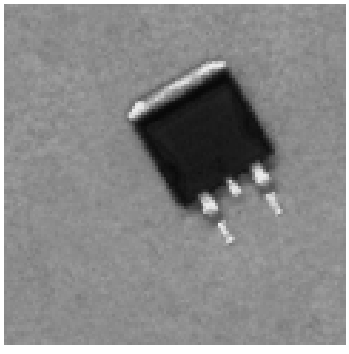

Figure 2. Reference image of the workpiece with defined translation (in this example 15 pixel to the right, 5 pixel up) and rotation angle (30.0°)

4. CNN APPROACH

With CNN technique two different methods have been realised. Due to the fact that CNNs are predestined for classification, specific rotation classes have been generated from the reference image set for training and pose estimation. In a second, more sophisticated method the system learns the parameters of an affine transformation – including the sought-after parameters of translation and rotation – by means of a regression process.

4.1 Classification method

As described in section 3 the image database includes several images of the same rotation angle. Therefore, 360 rotation classes (1° class width) and 180 rotation classes (2° class width) resp. could be created. The majority (6000 images) of the reference dataset has been used for the training of the CNN, the rest (3000 images) for the investigation of the extraction quality. The following aspects have been investigated:

- Number of iterations for the training
- Influence of "drop out" technique
- Influence of the number of rotation classes
- Influence of the translations

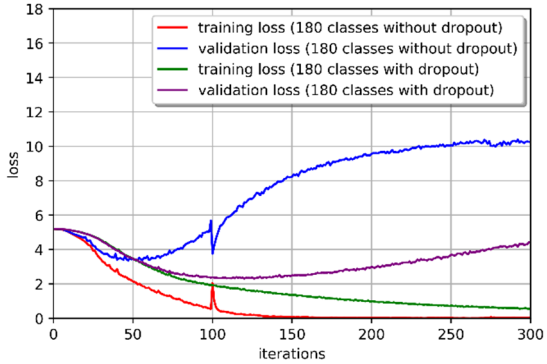

Figure 3. Training and validation loss dependent on the number of iterations

For the training of the CNN in the context of the classification approach 300 iterations have proved to be a sufficient number, more iterations introduce overfitting and the results get even worse. To reduce overfitting the so-called "drop out" technique had been introduced, i.e. the varying exclusion of several nodes (e.g. 50%) in the layers of the CNN for each iteration. Especially the validation loss can be significantly improved (Figure 3).

4.1.1 Results of classification method

The main problem of the pose estimation carried out by the classification method are a certain number of outliers of approx. 180° caused by the moderate symmetry of the workpiece. In Table 1 the number of outliers and the standard deviation σ of the determination of the rotation angle including and excluding these gross errors is given.

Another drawback is the relatively low accuracy of about ±4° to ±5° which can be obtained at best (Table 1).

# classes	# iterations	σ_α (w.o.)	σ_α (n.o.)	# outliers
180	300 (n.d.)	± 44.2°	± 4.9°	189
180	300 (w.d.)	± 22.9°	± 3.8°	49
360	300 (w.d.)	± 35.5°	± 4.3°	121

Table 1. Results of classification method: accuracy of rotation angle σ_α and number of outliers dependent on the number of classes, the determination with (w.o.) or without outliers (n.o.) and the use of drop out techniques (w.d.) or without (n.d.)

In specific investigations it could be verified that this is caused by the additional translations which have a significant influence on the accuracy. Tests applied to images without translations show much better results, especially concerning outliers which do not occur using these images (Table 2).

# classes	# iterations	σ_α	# outliers
180	300	± 0.8°	0
180	300 (w.d.)	± 0.8°	0
360	300	± 0.7°	0

Table 2. Results of classification method: accuracy of the rotation angle σ_α based on images without translations of the workpiece, dependent on determination with drop out techniques (w.d.) or without

4.2 Regression method

As described above the pose of a workpiece can be estimated alternatively by the determination of the parameters of an affine transformation based on a regression. Due to the fact that a discontinuity occurs at the transition of the rotation angle from 360° to 0°, the angle was defined as a complex number: $z = a + ib$ with $a=r\cdot\cos(\alpha)$ and $b=r\cdot\sin(\alpha)$. Therefore, in total four parameters have to be estimated instead of three.

4.2.1 Results of regression method

In a first project phase only images without different perspective views and illuminations had been used to remain comparable with the results of the classification method. Here again the necessary number of iterations for the training of the CNN is an important aspect. Our research has shown that more than approximately 6000 iterations will not lead to an improvement of the results. In Table 3 the influence of the number of iterations is shown.

To investigate the influence of the input images for training of the CNN on the results the number of basic images had been reduced to 10 from which 3000 and 6000 reference images resp. had been generated by rotation and translation. The effect of the fewer basic images is only marginal, while the reduced total number of training images (3000) led to a lower accuracy of the rotation angle (Table 4).

# iterations	σ_α	σ_x [pixel]	σ_y [pixel]
100	± 5.8°	± 1.1	± 1.1
300	± 3.6°	± 0.9	± 0.9
1000	± 2.8°	± 0.6	± 0.7
3000	± 2.4°	± 0.5	± 0.6
6000	± 2.1°	± 0.5	± 0.5
9000	± 2.2°	± 0.5	± 0.5

Table 3. Influence of the number of iterations on the accuracy of rotation angle σ_α and translation σ_x , σ_y

# reference images	σ_α	σ_x [pixel]	σ_y [pixel]
3000	± 2.9°	± 0.6	± 0.6
6000	± 2.1°	± 0.5	± 0.5

Table 4. Influence of a reduced number of reference images on the accuracy of rotation angle σ_α and translation σ_x, σ_y

For comparison with the classification approach also reference images without translations had been processed with the regression method (cf. Table 2). The accuracy of the rotation angle could be improved to $\sigma_\alpha = \pm 0.6°$.

Due to practical relevance one of the most interesting investigations was the processing of reference images with different perspective views and illuminations of the workpiece. As could be expected, Table 5 confirms to degradation of the accuracies with the largest effect on the determination of the rotation angle.

# iterations	σ_α	σ_x [pixel]	σ_y [pixel]
6000	± 3.5°	± 0.7	± 0.7

Table 5. Influence of different perspective views and illuminations of the workpiece on the accuracy of rotation angle σ_α and translation σ_x, σ_y

5. CLASSICAL IMAGE PROCESSING APPROACH

To compare the results of the two CNN based methods described above a classical image processing approach had been applied to the same dataset of reference images. The well-known commercial HALCON system (Halcon, 2018) had been used for this purpose. The detection and pose estimation traditionally is based on edge detection algorithms, in the case of HALCON a so-called shape model has to be created (Figure 4). This had been carried out with one of the basic images of the transistor element without rotation and translation. Essentially, the value for the contrast of significant edges has to be defined to obtain a sufficient model. If it is too low, too many weak edges are extracted, if it is too high, too few edges can be detected for the model which may lead to a distinct incompleteness. Due to the fact that the origin of this shape model differs from that of our CNN approach, an offset has to be regarded especially for images with rotated workpiece.

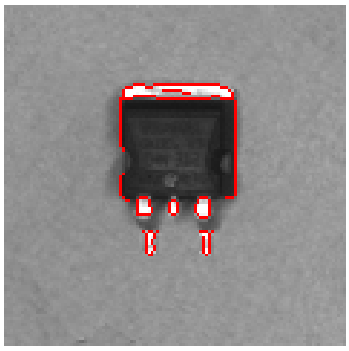

Figure 4. Shape model of the transistor element created by HALCON system

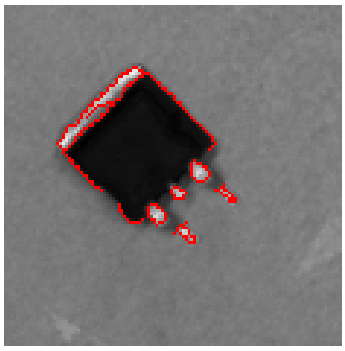

Figure 5. Detected workpiece in a reference image and its estimated pose by HALCON system

To determine its quality this approach had been applied to a subset (70 images) of the reference dataset already described in section 3, because only the shape model has to be created from one of these pictures instead of a training with thousands of images.

5.1 Results of classical image processing

To obtain sufficient results with HALCON’s shape model described above several parameter values have to be set by the user. The most important ones are the score value (a measure (0 to 1) of the accordance between shape model and extracted object model) and the minimum contrast (threshold for detection of edges of an object to be extracted). For their specification a test series with varying values has to be performed during the preparation phase. Table 6 shows insufficient parameter values producing a number of missing detections as well as appropriate settings for faultless results. If these parameters are set to appropriate values (e.g. min. contrast = 0.6, score = 0.4) a detection rate of 100% with no outliers can be reached (cf. Figure 5) as well as comparable accuracies of rotation angle and translations (Tables 7 and 8).

min. contrast	score	undetected
0.75	0.65	53 %
0.70	0.65	31 %
0.60	0.40	0 %

Table 6. Different parameter setting for model extraction in HALCON. With appropriate values a detection rate of 100% can be obtained

In a first step the generated shape model was applied only to images of our workpiece without perspective effects (Table 7), i.e. the workpiece was positioned in the middle of the images.

σ_α	σ_x	σ_y
± 0.4°	± 0.2 pixel ± 0.04 mm	± 0.5 pixel ± 0.13 mm

Table 7. Accuracy of rotation angle σ_α and translation σ_x, σ_y obtained by processing reference images without perspective effects

Of course, more realistic for industrial applications are situations where different perspective views and different illumination effects occur (Figure 5). In Table 8 the results for these images of our dataset are assembled. As expected the accuracies decrease a little bit but not in a drastic way.

σ_α	σ_x	σ_y
± 0.5°	± 0.4 pixel ± 0.11 mm	± 0.4 pixel ± 0.11 mm

Table 8. Accuracy of rotation angle σ_α and translation σ_x, σ_y obtained by processing reference images with perspective and illumination effects

It can be stated that with this classical approach of image processing a good quality and robustness of the extracted pose parameters can be reached which fulfil most requirements of industrial applications. Especially the accuracy of rotation angle is much better than in other methods.

6. DISCUSSION

The classification approach by CNNs seems to be no sufficient method for detection and pose estimation in this application. The main drawback is the problem of determining the translations of the workpiece and unknown translations lead to decreasing accuracy of the rotation angle. Only if the object is rotated exclusively (without translations) the results for the rotation angle are of good quality (but it seems to be no realistic condition). Additionally, too many errors occur due to a significant number of outliers caused by the moderate symmetry of the workpiece. Acceptable results with CNNs can be

reached by the regression method. The complete 2D pose can be obtained without outliers. The accuracy of the rotation angle is about $\sigma_\alpha = \pm 2°$ without different perspective views and illuminations, but again, this is not a realistic scenario in the practice of industrial production. Taking these influences into account the quality of approx. $\sigma_\alpha = \pm 3.5°$ can be rated only as medium level. In contrast, the accuracy of translations (approx. $\sigma_{x,y} = \pm 0.7$ to ± 0.8 pixel) is comparable to classical image processing algorithms. As disadvantage the significantly longer processing time – compared to classification and classical image processing – must be mentioned. In general, the relatively long time for training of the CNN may be another important aspect (dependent on the application).

In comparison classical image processing algorithms deliver good results with high reliability concerning detection rate and number of outliers as well as a high accuracy of the extracted pose parameters. A detection rate of 100% with no outliers could be obtained, the accuracy of the rotation angle was about $\sigma_\alpha = \pm 0.5°$ (approx. factor 7 better than CNN approach) and about $\sigma_{x,y} = \pm 0.4$ pixel for the translations. But it has to be taken into account that for this approach also a preparation phase (instead of a training phase like for CNNs) has to be carried out to find appropriate parameter settings (e.g. for contrast, score etc.), i.e. the parameter values have to be adapted for each new object and its environmental conditions. On the other hand the user can influence the resulting accuracy: dependent on the setting of these parameters the quality of – for instance – the rotation angle can be improved accepting a loss of translation accuracy.

If CNN regression approach should be compared to classical methods of image processing it can be stated for this application, that CNNs show some disadvantages: first of all the quality of rotation angle determination has to be mentioned which – in our project – was worse about a factor of 4 while the accuracy of the translations lay in the same dimension. Furthermore, the amount of training data is much higher where hundreds or thousands of reference images have to be generated. Finally, time for the training of CNNs can be enormous, in our application approx. 10h for 6000 iterations (processor: 3 GHz, RAM: 16 GB). In contrast up-to-date image processing algorithms are fully developed and produce results of high accuracy and reliability in short processing times. Therefore, these systems may be preferable for similar tasks in industrial domain.

7. CONCLUSION

In this contribution two CNN approaches have been created for 2D pose estimation of industrial workpieces. For comparison a classical image processing system had been applied to the same reference images. Exemplarily a transistor element was used for these investigations. It can be concluded that the classification approach of CNNs cannot fulfil the requirements of pose estimation while determining the parameters of an affine transformation by regression is able to extract such workpieces with a moderate accuracy and reliability of the rotation angle, the values for the translations are even comparable to classical approaches of up-to-date image processing algorithms. The necessary time for the training of the CNN (including the generation of an image database with a sufficient number of images) has to be taken into account. It can reach enormous dimensions and is typically much longer than e.g. creating a shape model (e.g. approx. 10h in our application). CNNs have undoubtedly advantages, if the distinguishable features and characteristics of an object are vague or unclear. For pre-defined workpieces in industrial environment classical image processing systems seem to be predominant in terms of quality and performance.

ACKNOWLEDGEMENTS

The authors would like to thank Dipl.-Inf. Alexander Piaseczki (SAC Sirius Advanced Cybernetics) for his comprehensive support of this research project.

REFERENCES

Bourez, C., 2017. *Deep Learning with Theano*. Packt Publishing, 2017, ISBN 978–1–786–46305–0

Géron, A., 2017. *Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme*. O'Reilly, Heidelberg, 2017, 1st edition, ISBN 978–3–96009–061–8

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep learning*. The MIT Press, Cambridge, Massachusetts, 2016 (Adaptive computation and machine learning), ISBN 978–0–262–03561–3

Halcon, 2018. http://download.mvtec.com/halcon-6.1-reference-hdevelop-de.pdf (8 May 2018)

Lasagne, 2015. http://lasagne.readthedocs.io/en/latest/modules/objectives.html (5 January 2018)

Patterson, J., Gibson, A., 2017. *Deep learning: a practitioner's approach*. O'Reilly Media, Sebastopol, 2017, 1st edition,
ISBN 978–1–491–91425–09781491914250