Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

Proposal

Lecture Computer Vision WS 20/21

Bottle Cap Detection by Lokalizing Regions of Interest

by

Janelle Pfeifer

Matr.Nr: 9030898

Submitted on: 9. Dezember 2020

1 Introduction

The topic of this proposal is to describe a computer vision approach that finds bottle caps in a set of videos. The goal is to detect and locate metal bottle caps in static scenes within videos. The bottle caps are inside of a container with a homogenous background. Each bottle cap can be in either of three states: it can be face-up, facedown, or deformed. The state the bottle cap is in needs to be determined. Together with the bottle caps, there are distractor objects in the container. These distractors can have a similar appearance to the bottle caps.

Finding the bottle caps and their states mirrors a visual quality control task in which manufactured objects get inspected via a camera to look for manufacturing errors. When it comes to bottle caps specifically, a factory needs to make sure that there are no bottles without a cap, no caps are damaged and all caps are correctly placed on their bottle. This is similar to our task. A deformed bottle cap is a damaged bottle cap in quality control. A face-down bottle cap is not placed correctly and a distractor object is the bottleneck of a bottle that didn't get a cap at all.

In their Article Prabuwono et al. (2019) propose a solution for detecting plastic bottle caps that are not placed correctly. They use a conveyor belt to move bottles by a camera one by one. Additionally, they use an led for lighting and a black box behind the bottles to reduce noise. In an industrial context, the environment is very controlled. The lighting and the position of the bottle are always fixed. T

2 Material and methods

The videos are 10 to 20 seconds long. They have a frame rate of 25 fps and are compressed as mp4 files. They start by showing a rectangular container in a static scene. Next bottle caps and other 'distractor' objects get placed inside the container. The container with the objects is shown in a static scene afterward. Lastly, the objects get removed from the container. Bottle caps that are outside of the container can be ignored. At the start of the video, there are no bottle caps inside the container. Distractor Objekts however may be present and they may remain in the container at the end of the video. There are between 3 and 15 bottle caps inside the container.

Bottle caps can be present in three different states: face-up, face-down, or visibly deformed.

- When a cap is face-up it most likely has a colorful motif. Caps can have any color and they have a circular shape.
- When a cap is face-down it is white with a visible zig-zag rim. The cap has a circular shape.
- a deformed cap is either white or colorful and it doesn't have the circular shape of face-up or face-down caps and it is slightly smaller.

2.1 Approach

From the data description, it can be assumed that every video has two static scenes. One at the start of the video and one closer to the middle. Both scenes show the same container. Once without any bottle caps and once with bottle caps. Comparing the scenes with one another will provide a good indication of regions where a bottle cap might be. Since there are no bottle caps in the first static scene, every region that is different in the first and the second scene can contain a bottle cap. Furthermore, these regions are the only areas of the scene where bottle caps can be. To compare the static scenes, they need to get found within the video and for both of the scenes, one representative frame needs to be chosen.

To find the static scenes the differences between consecutive frames get measured. Multiple similar frames indicate a static scene. Andy and Haikal (2017) search for duplicate frames in videos to detect video forgery. They calculate the hash value of each frame and compare them. This approach finds frames that are exactly the same. However, the frames of static scenes are not exact duplicates due to noise, camera shake, and compression.(Andy and Haikal 2017; ryanfox 2015)

Norollah et al. (2012) use template matching to compare X-ray dental images. By using template matching they can correct differences like rotation, scale, and noise between the images. When looking for similar frames template matching is invariant to noise and camera shake. Because of this template matching is used to quantify the differences between video frames to find static scenes. (Norollah et al. 2012)

There is no guarantee that the videos do not contain more than two sequences. To identify the relevant static scenes the two sequences that have the longest duration (the largest amount of frames) and have the most fitting time frames in the video get chosen. To decrease the number of frames that need to be compared to one another the knowledge that one of the scenes should be at the start and one should be near the middle can be used. If there is a sufficiently long sequence at the start of the video it may be immediately assumed that the first static scene is found. For the second static scene, the frames in the middle of the video may be tested for a sequence. If a sufficiently long sequence is found there is no need to search for more sequences. One frame for the two static scenes gets chosen as representative. Since the frames are all very similar it is irrelevant which frame is picked.

The frames of the first and second scene get compared to find regions in which the frames differ from one another. Those are regions of interest since they can contain bottle caps. To find these regions the images are subtracted from one another. Then a morphological close operation is run to remove noise and to better distinguish regions that are close to one another.

To further reduce the regions of interest the rectangular container with the homogenous background is found in the frames. Regions of interest that are outside of the container are no longer interesting. Even if they contain a bottle cap those should be ignored.

Next, the contents of the regions of interest get analyzed. For each region, the corresponding section of the frame from the second static scene is cropped out. This cropped out section gets analyzed by a TensorFlow machine learning model. The model is trained for object recognition with cropped regions from a set of training videos to detect bottle caps that are face-down, face-up, and deformed. The model determines for every cropped image if it contains a bottle cap and what kind.(Jiao et al. 2019)

The found bottlecaps and their type are then illustrated in the frame from the second scene.

3 Anticipated results

Finding regions of interest is expected to be quite robust as long as lighting conditions stay constant within individual videos. If the container as a whole gets moved or changes its position in the frame it could be difficult to consistently find the regions of interest. To fix this possible issue the container can get detected in the frames from both scenes. When the positions of the containers get matched to one another the differences between the frames can get analyzed correctly.

Recognizing the bottle caps that are facing down with TensorFlow is expected to work well. The bottle caps are nearly always white with a zig-zag outline when they are face-down. A machine learning model should be able to distinguish this from other objects. Face-up bottle caps might be harder to detect and deformed bottlecaps might get confused with distractor objects because they can be colorful and without the distinctive round shape.

Changing lighting conditions or hard shadows could present a problem to the detection approach because they get picked up as regions of interest as well. However, the machine learning model should dismiss them.

The approach for finding regions of interest gets tested with videos that have been taken by students. The TensorFlow model gets trained with images that have been cropped from those videos. The model will get tested with a different batch of videos.

4 Literaturverzeichnis

S. Andy and A. Haikal. Simple duplicate frame detection of mjpeg codec for video forensic. In *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pages 321–324, 2017. doi: 10.1109/ICITISEE.2017.8285520.

L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu. A survey of deep learning-based object detection. *IEEE Access*, 7:128837–128868, 2019.

M. Norollah, H. Pourghassem, and H. Mahdavi-Nasab. Image registration using template matching and similarity measures for dental radiograph. In *2012 Fourth International Conference on Computational Intelligence and Communication Networks*, pages 331–335, 2012. doi: 10.1109/CICN.2012.129.

A. S. Prabuwono, W. Usino, L. Yazdi, A. H. Basori, A. Bramantoro, I. Syamsuddin, A. Yunianta, K. Allehaibi, and S. Allehaibi. Automated visual inspection for bottle caps using fuzzy logic. *TEM Journal*, 8:107–112, 02 2019. doi: 10.18421/TEM81-15.

ryanfox. retread, 2015. URL `https://github.com/ryanfox/retread`. [Online; Stand 25. November 2020].