SHORT TERM SCIENTIFIC MISSION (STSM) – SCIENTIFIC REPORT

The STSM applicant submits this report for approval to the STSM coordinator

Action number: Action IC1307 - 39993
STSM title: Describing Ear Images Using Natural Language
STSM start and end date: 05/02/2018 to 16/02/2018
Grantee name: Žiga Emeršič

PURPOSE OF THE STSM/

(max.500 words)

The purpose of the STSM was to develop a system that will not only remove/replace ear accessories, but also describe the decisions it made in a natural human language through the following steps:

- Ear accessories pixel-wise detection and segmentation using state-of-the-art approaches.
- Replacement of ear accessories with appropriate content that looks like an ear.
- Description in natural language of what was done – if and how the images were modified.

This is planned, not only to increase the overall robustness of ear recognition, but also to increase the transparency of the procedure to the experts (e.g. forensic experts) using the system. Our goal was to prepare a showcase of such a system that would do these steps. We investigated possible solutions and also implemented initial solutions. Still there is a lot to do and room for improvement, but we present some preliminary, qualitative results to show the feasibility of our approach. Steps of segmentation and infilling are not yet connected, because additional training in both parts is still required. Nevertheless, with the preliminary results we have proved that the whole pipeline is feasible and close to the final implementation.

Ear accessories and occlusions in general are one of the most problematic aspects of ear recognition [1]. There are two underlying reasons for this:

- Accessories and other occluding objects (such as hair) occlude areas that carry biometric information.
- Accessories themselves carry certain identifying features: e.g. a person carries same earrings in all images used for enrolment. The decision model used for ear recognition can then implicitly use information from earrings (and learn based on features obtained from accessories). This can lead to two types of errors: (1) person does not wear that type of ear accessories in probe images and the decision model fails to recognize that person, or (2) a person wears a type of earrings used by some other person during the enrolment stage and gets recognized as that person.

To counter these issues, we need to detect ear accessories (and other objects that occlude ears) and then replace them. Here, however, there are multiple options: replacing them with certain color (black, white, alpha channel, etc.). Although this circumvents problem (2), because we remove identifiable features, there

still remains problem (1). Furthermore, neural network can erroneously presume some identity traits from the shape of a black/white area.

To counter these two issues, we therefore need to remove or replace earrings. However, if do that we need to be aware that such actions happened. This especially holds true for forensics applications where experts need to be aware of any tinkering of the original images. Presenting evidence in court and deciding on data can be problematic if the decision-makers are not aware of all the steps in the process.

Therefore it is necessary to prepare a description of unwanted, non-biometric objects that were removed or replaced and present them in a readable form for the human in the loop.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

(max.500 words)

We focused on infilling accessories with the natural looking textures and providing descriptions of what was done. In order to achieve that we went through the four steps described below.

STEP 1: DATASET PREPARATION

Convolutional neural networks require large amounts of data to successfully train a specific problem domain. This poses a problem when trying to train a network to detect ear accessories, because there are to the best of our knowledge no datasets available that would contain images of ears with pixel-wise annotated accessories. We solved this problem by manually preparing a dataset with artificially pasted images of 10 earrings into the 1000 images of ears of AWE dataset [1]. This means that we also trivially acquired pixel-wise annotations in a form of masks of earrings. However, in order to prepare diverse and sufficient amount of data we randomly positioned earrings and randomly changed colors.

SETP 2: PIXEL-WISE DETECTION & SEGMENTATION OF EAR ACCESSORIES

In this step we prepared two setups: modified SegNet architecture [2, 3] and Refinet [4]. The reason is that the latter presents a state-of-the-art in the domain of pixel-wise annotation, and SegNet used to be an important benchmark and its performance can be more easily compared to previous work. In case of SegNet we used a Caffe implementation with parameters set to default. In the case of Refinenet we used original parameters and used original implementation in Matlab provided by the authors.

STEP 3: REMOVAL OF ACCESSORIES BY TEXUTRE INFILLING

For the accessories removal we decided to use the so called in painting approach based on CNNs [5]. The reason is that this approach gives the most realistic results and thus (we presume) is the most perspective when ear recognition is considered.

Figure 1: A sample of a segmented, ear accessory (artificially generated) and an overlay with the location map resulting in a description. Location map is actually a matrix with values: 1, 2, 4, 8, 16 visualized here with grayscale tones.

STEP 4: DESCRIBING DETECTED ACCESSORIES WITH NATURAL LANGUAGE

For the final step we decided to use map overlay of predefined regions over the detected areas of accessories as shown in the following Figure 1.

The description is composed using the following rules. All images are resized to a uniform size, but only so that the logic for preparation of description is simpler; images are not stored afterwards. Each detected region is observed, but areas with surface area smaller than 30 pixels (e.g. 5 x 5 area would be discarded):

- Size description: small: 30px < 1000px, medium: 1000px < 3000px, large: > 3000px.
- Location description: area is overlaid over the location mask as shown in Figure 1. Both matrices are multiplied and then 'unique' operation is applied. We are left only with values where blob was present. Because mask is represented with "binary" numbers we can sum these unique number and create simple if-then rules. E.g. 1 means center, 2,3,4,5 mean top, 2, 3, 8, 9 left and so on. E.g. 21 would mean that blob stretches from top to bottom and location description is omitted in such case.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

(max. 500 words)

The core results that were the main goal of this STSM are satisfactory, this means that the resulting descriptions themselves work well for the domain. However, each separate step still needs work and improvement. This means that we have not yet connected the whole pipeline into one, but used separate inputs into each step, because steps prior to each step did not provide data useful enough to facilitate the next steps.

For example, during the detection step we get a lot of miss-classifications of ear accessories, although we achieved good results when we segmented ears from face images in our previous work [2]. The results for pixel-wise detection and segmentation are therefore not satisfactory yet, but show great promise. But, most importantly, the preliminary results did not expose any issues that would prevent good accessories detection, only more time is needed for CNN training, thus improving results and joining up the whole pipeline.

Figure 2:Two samples of preliminary infilling results on artificially generated black occlusion boxes.

In the step of textures infilling we still have problems when smaller areas are used (the ones similar to earrings). However, the preliminary results using larger, artificially generated squares look promising, as shown in Figure 2. More training is needed, however, to achieve useful results on smaller objects.

The description step uses input from the first step (accessories detection), so we used artificial input here. The description logic works well with an example shown in Figure 1.

We expect two major contributions from this work: (1) With the segmented (either removed or replaced) ear accessories method we plan to enable future researchers to increase robustness of their ear recognition pipelines. (2) With the natural-language-described removed/replaced objects we plan to increase the

transparency of ear recognition processes and make the decisions made more readable, not only for the experts but also for more general users, present in the e.g. forensics process.

FUTURE COLLABORATIONS (if applicable)

(max.500 words)

Future collaboration would mainly tackle the area of object recognition. Based on the pipeline outlined and developed during this STSM the final goal is to improve ear recognition approach. However, because in the process of ear recognition we plan to use general-purpose convolutional neural networks for classification tasks, these networks can easily be applied to other domains.

The research team in Las Palmas, under the leadership of professor Carlos M. Travieso-González is dealing with many object classification tasks, currently, one of the problems is the tree bark recognition. Our plan is to share models and knowledge between the teams, thus improving results on both ends. Furthermore, they are also active in the biometric domain - we also co-authored a paper together with professor Miguel A. Ferrer on the topic of sclera segmentation, so this is another possible domain of future collaboration with the University of Las Palmas.

REFERENCES

[1] Ž. Emeršič, V. Štruc, and P. Peer, “Ear Recognition: More Than a Survey,” Neurocomputing, 2017.

[2] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, “Convolutional Encoder-Decoder Networks for Pixel-Wise Ear Detection and Segmentation,” IET Biometrics, 2018.

[3] Badrinarayanan, V., Handa, A., Cipolla, R.: ‘SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling’, arXiv:150507293, 2015

[4] Lin, Guosheng, et al. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[5] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.