Intelligent embedded systems for facial soft biometrics in social robotics

Abstract – Vincenzo Vigilante

The foundation of this thesis is the observation of the usefulness of contextual clues in the context of social robotics: studies prove that human-like behaviour is key to generate in the interlocutor the feeling of empathy that allows him to subconsciously perceive the robot as his peer.

From the analysis of the faces of the people around, the robot can gather information that allows to personalize the interaction and enhance the feeling of empathy given by the robot. Such information, including age, gender, ethnicity, emotion and more is called "soft biometrics" because it does not allow unique, perfect, identification of a person, but it is nevertheless used by humans to distinguish their peers;

We observe that tasks in the domain of facial soft biometrics are extensively studied in literature but the application to realistic conditions such as the ones of social robotics, introduces some constraints that require specific attention, namely resource constraints and robustness constraints. Resource constraints are limitations due to the actual hardware that runs the prediction systems; such constraints for example require the memory footprint to be confined to what the hardware can handle and require the inference time to be limited as well, in order for the information to be available in good time to be used in a naturally paced iteration. Robustness concerns the ability of the system to produce correct predictions based on input images that are affected by all kinds of corruptions and perturbations that are present on images acquired in unconstrained conditions using typical hardware from the considered application; for instance, embedded cameras produce noisy images with limited resolution and dynamic range.

In this thesis we tackle those themes in the context of Deep Learning. We design and evaluate efficient and effective CNN-based methods for the tasks of gender recognition, ethnicity recognition, age estimation and emotion classification.

For gender recognition we observe that traditional CNN architectures are designed with reference to the problem of object recognition and evaluated on the ImageNet benchmark. We argue that the task of gender recognition has different characteristics, therefore we design an efficient architecture based on the MobileNet v2 architecture but with reduced depth, input size and number of feature maps. We experiment our reduced architecture on the LFW+ public benchmark and we compare with the state of the art. We show that our architecture is capable of recognizing gender with an accuracy of 98.1% in just 56ms on an embedded device without any neural network acceleration, which we judge perfectly reasonable for the application to social robots.

When considering the state of the art of Ethnicity recognition, we conclude that the development of effective methodologies is held back by the absence of a large dataset. We effectively design a dataset that is large (3.3 million images, 9000 different identities) and reliably annotated, by having multiple people of different ethnicities participate in the annotation of the same data. We argue that training on our dataset makes neural networks more accurate than training on other datasets; we prove this by training different neural networks, including the efficient architecture MobileNet v2 cited before, and using an independent benchmark to assess that the accuracy is indeed higher. We believe that our results have great margins for improvement, in fact we consider those as a baseline, and make our dataset (VMER) publicly available to foster further development of the state of the art on this task.

With respect to age recognition, when trying to train a fast architecture for this task (as we did for gender and ethnicity) we find that the publicly available datasets are either too small or contain a very noisy automatic annotation; this is understandable, since annotating a dataset containing millions of images is an extremely costly task. Existing accurate methods that have been proposed until now, resort to the use of costly manual cleaning procedures, and use large ensembles of neural networks which are better resilient to

the inadequacy of the dataset than simpler architectures. Those ensembles are of course slow and large and inadequate for practical applications; for instance, the winner of the prestigious LAP 2016 challenge takes about 6 second per image and requires the use of a powerful GPU: given the constraints of our application, such a solution is not viable for us. We propose the use of a technique called knowledge distillation to overcome the issue: we use the powerful but slow ensemble that we just described (that we call teacher) to annotate a very large dataset (that we call VMAGE), and then we use that dataset to train simple architectures (students). We show that our students consistently outperform the same architectures trained with the commonly used dataset from the state of the art, without the use of the distillation technique. This proves the effectiveness of our approach. Additionally, we compare the accuracy of our benchmark on all the major public benchmarks and prove that, under different protocols, our architectures achieve competitive accuracy with respect to existing literature, while being simpler and more efficient, thus adequate for use in our proposed application.

Finally, with reference to the task of Emotion Recognition, we experiment the effect of different image corruptions and perturbations from the real world on 4 different architectures (VGG, SENet, DenseNet and Xception). We evaluate the effect of Autoaugment and antialiased downsampling on those architectures, the first being a technique for effectively augmenting the training data and the second being an architectural modification that adds low pass filters inside the networks wherever a downsampling happen (e.g. max-pooling or strided convolutions). For our evaluation, we construct a benchmark data set on top of the RAF-DB test set that includes images with corruptions that typically occur when the recognition systems are deployed in real scenarios. Corruptions include different kinds of blur (motion blur, lens blur, zoom blur, gaussian blur), of noise (gaussian noise, shot noise), pixelation, jpeg compression, changes in brightness and contrast and combination of those. For evaluating the stability of the predictions, we generate the RAF-DB-P dataset, that includes versions of the testing images where we perturb the brightness, the position, the scale, the rotation, the quantity of blur and the pattern of noise. We find that the combined use of antialiasing and Autoaugment substantially contributes to the improvement of the robustness to corruptions, especially to those of the noise and digital type, of SENet and DenseNet. The VGG architecture instead showed the highest classification stability with respect to perturbations that affect subsequent frames of a sequence, especially when combined with the use anti-aliasing filters. We find the Xception methods to be not suitable for facial emotion analysis in our setting, since they are expecially affected by corruptions and perturbations. In conclusion our experiments demonstrated that the common corruptions and perturbations are important aspects to take into account when evaluating methods to be deployed in real scenarios. However, none of the existing methods, which we modified with anti-aliasing filters and trained with extensive data augmentation, showed robustness to all the considered corruptions and perturbations, thus this aspect remains open for future investigation.

Overall, in this work, we were able, for each of the four tasks, gender, age, ethnicity, emotion, to design a CNN-based system able to achieve state of art performance while being able to perform in the target social robotic environment, with limited inference time and memory requirements and able to work in reasonably "wild", uncontrolled settings. Future work needs to focus on the problem of robustness and propose a solution for a more reliable perception system.