Real-Time Age Estimation from Facial Images Using YOLO and EfficientNet

Chapter · October 2021

DOI: 10.1007/978-3-030-89131-2_25

Real-Time Age Estimation From Facial Images Using YOLO and EfficientNet

Giovanna Castellano[0000−0002−6489−8628],
Berardina De Carolis[0000−0002−2689−137X], Nicola Marvulli,
Mauro Sciancalepore, and Gennaro Vessio[0000−0002−0883−2691]

Department of Computer Science, University of Bari "Aldo Moro", Italy
{giovanna.castellano,berardina.decarolis,gennaro.vessio}@uniba.it
{n.marvulli1,m.sciancalepore20}@studenti.uniba.it

Abstract. Automatic age estimation from facial images is attracting increasing interest due to its many potential applications. Several deep learning-based methods have been proposed to tackle this task; however, they usually require prohibitive resources to run in real-time. In this work, we propose a fully automated system based on YOLOv5 and EfficientNet to perform face detection and subsequent age estimation in real-time. Also, to make the model more robust, EfficientNet was trained on the new MIVIA Age Dataset, released as part of a challenge. The results obtained in the contest are promising, and are strengthened by the lightness of the overall system which in fact is not only effective but also efficient.

Keywords: Computer Vision · Age estimation · Convolutional neural networks · Face detection.

1 Introduction

Age estimation aims to predict a person's age by analyzing face images. It is even a difficult task for humans and, as such, it has always been a challenging problem in Computer Vision [4, 22]. An accurate age estimate from facial images would be beneficial for several real-world applications. For example, it can be useful in e-commerce applications to automatically suggest advertisements according to people's ages and show the content they are most likely to like, thus improving engagement [9]. Furthermore, in the field of safety and security, and in particular of human-computer interaction based on age, such systems can be used to prevent minors from browsing sensitive content [21]. Age estimation systems also prove effective when implemented on social robots, which typically take advantage of soft biometrics, including age, to show an empathic behavior. This is desirable in all contexts that require a natural interaction with humans, such as elderly care [2], museum guide [5], and so on.

The countless number of applications has brought great attention to age estimation systems; in fact, research in this field is very active both for the interest in its applications, but also for the great challenges that arise when trying to solve this type of task [1]. The main challenges are related to the quality of the facial image, the different poses, orientations, occlusions, bad light conditions, and the presence of facial gadgets. The classical approaches to solving these problems are based on the manual extraction of features and generally consist of two different stages: one dedicated to the extraction of invariant and robust features that encode information on aging; the other dedicated to age estimation, using traditional machine learning algorithms (e.g., [11, 13]). The disadvantages of the manual approach lie in the fact that the feature extraction process is closely related to the human experience and also requires a lot of effort. For these reasons, deep learning-based solutions for age estimation have begun to spread, thanks to their ability to build end-to-end age estimation systems, eliminating the need for any kind of manual feature engineering (e.g., [3, 20]).

In this regard, it is worth underlining that the most promising methods proposed in the literature use complex architectures or ensembles of deep models, whose resulting classifiers are not usable in real applications, as they require prohibitive computational resources that are not always available. Furthermore, their training procedure, made even more complex by the plurality of neural networks to train, typically requires huge training sets that are not easy to collect. To address these issues, in this work we use a huge data set recently released, that is the MIVIA Age Dataset [10], and propose a method based on a single neural network, i.e. EfficientNet [26], to provide an effective and at the same time efficient model for age estimation suitable for real-time applications. In addition, to fully exploit the potential of the proposed model, it is combined with the lightweight YOLOv5 detector [15] to develop a fully automated face detection and age estimation framework.

This article accompanies our participation in the "Guess the Age" contest,[1] as part of CAIP 2021, in which the MIVIA Age Dataset was released and whose goal was to specifically design single network solutions.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 describes the proposed method. Section 4 presents and discusses the experimental results. Section 5 concludes the paper.

2 Related Work

Early work attempting to solve the age estimation problem focused primarily on building robust sets of aging features, such as facial features and wrinkles [16], facial aging patterns [8] and biologically inspired features [13], taking into account different orientations and scales. General texture description features have also been exploited, such as Local Binary Patterns [11] and Gabor features [7]. Given the features extracted, age prediction is performed with a chosen learning algorithm, such as Support Vector Machines, Linear Discriminant Analysis, and so on. Further studies found that other traits, such as gender and ethnicity [6, 12], can improve the overall performance. Age estimation can also be treated as a classification task rather than a regression task [17].

[1] http://gta2021.unisa.it/

More recent works have begun to adopt techniques based on deep learning, some of which exploited rather complex architectures, such as the multi-scale architecture proposed in [29] and the tree-structured architecture proposed in [18]. Malli et al. [3] proposed an ensemble-based solution, training different deep models and then averaging their results to get the predicted age. Othmani et al. [20] recently showed that fine-tuning a state-of-the-art model, such as Xception or VGG16, on large age-related datasets, can improve predictive accuracy. While effective, these solutions are not always able to meet the stringent computational requirements of most critical real-time applications. With the aim to meet real-time requirements, in this work we propose a combination of lightweight models to provide not only an effective, but above all an efficient solution for real-time face detection and subsequent age estimation.

3 Proposed Method

The general framework was developed as a Web App, written in Python. The server side is built on the Flask library, which receives the image stream from the client and responds with a JSON file that contains the coordinates of the bounding box of the detected face, the confidence score and the age estimate. The core of the system is a two-stage process based on two different deep learning models. In the first stage, the images/frames streamed by the client are fed into a face detector based on the YOLOv5 architecture to extract face ROIs. In the second stage, an age estimation model based on the EfficientNet architecture runs to predict the age. These two steps are described in more detail in the following subsections. It should be noted that the system can operate in different modes, depending on the user's needs. It can perform inferences in real-time using a webcam, or in batch mode, feeding it with a single image, video or a folder with many of them.

3.1 Face Detection

To implement the face detector, we used YOLO ("You only look once"). YOLO is a family of single-shot object detection models, which aims to surpass demanding region-based detectors, in order to create lightweight but solid object detection systems capable of running on mobile or edge devices providing accurate performance in real-time [23]. The main idea of YOLO is to represent each image as a grid of $S \times S$ cells. If the center of an object falls within a cell, that same cell is the responsible for the object. Each cell of the grid predicts B bounding boxes, assigning them confidence scores. Finally, non-maximum suppression is applied to clear overlapping or low-scoring boxes, given a certain threshold.

Among the different implementations of YOLO proposed in the literature, we adopted the latest Ultralytics implementation, namely YOLOv5 [15]. The adoption of YOLOv5, compared to other previous, yet powerful versions, is mainly due to its high versatility in the integration process and also to its lightweight structure that counts only 7.3M parameters. In fact, we have focused in particular on the lighter version, which is called YOLOv5s. The main difference between YOLOv5 and its predecessors is that it is implemented in PyTorch rather than being a fork of the original Darknet framework. In addition, some important improvements have been added, including self-learning bounding box anchors.

3.2 Age Estimation

To implement the age estimator, we used EfficientNet, a convolutional neural network that relies on AutoML and a so-called *compound coefficient* to uniformly scale its depth, width and resolution [26]. Unlike the conventional practice of arbitrary scaling these factors, in fact, the EfficientNet scaling method uniformly scales the width, depth and resolution of the network with a set of fixed scaling coefficients. The compound scaling method is motivated by the intuition that if the input image is large, the network needs more layers to increase the receptive fields and more channels to capture finer-grained patterns. EfficientNet currently achieves state-of-the-art performance across multiple benchmark datasets, but with an order of magnitude fewer parameters. The core EfficientNet-B0 network builds on the inverted residual blocks of MobileNetV2 [25], plus squeeze-and-excitation blocks. EfficientNet-B0 has been enhanced with the compound scaling method to achieve a family of models from B0 to B7. We chose EfficientNet-B0 to build our real-time age estimation model, as is the lightest, with 237 layers but only 5.3M parameters.

To adapt the model for the age estimation task, we removed the original classifier on top, replacing it with a global average pooling layer, a batch normalization layer, a dropout layer (with a small dropout rate of 0.2), and a single output neuron with linear activation.

4 Experiment

4.1 Datasets and Setting

YOLOv5 is a family of object detection architectures pre-trained on the COCO dataset. To fine-tune YOLOv5 for face detection, we used the Wider Face dataset [28] which includes 32,203 images, labeled in Pascal VOC format. We considered only a subset of 2200 images, which were selected excluding images with too small bounding boxes, i.e. in which the faces are not sufficiently large. To be used with YOLOv5, Pascal VOC labels (left, top, right, bottom) were converted to YOLO format (width, height, x center, y center). The dataset was split into a training set of 1600 images, a test set of 400 images and a validation set of 200 images. Fine-tuning was performed unfreezing the top 25 layers and using a learning rate of 10^{-2}, a mini-batch size of 32 and the standard YOLO loss function.

To train the age estimator, we considered the MIVIA Age Dataset which is the main dataset of the “Guess the Age” contest. It consists of 575,073 images

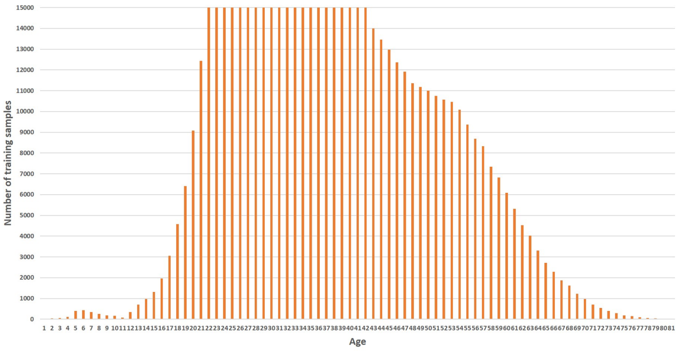

Fig. 1. MIVIA Age Dataset distribution by age (http://gta2021.unisa.it/).

of over 9,000 identities, obtained at different ages. It is worth mentioning that the MIVIA Age Dataset is among the largest publicly annotated datasets available on faces. Images were extracted from the VGGFace2 dataset and annotated with age using a "knowledge distillation" technique, making the dataset heterogeneous in terms of face size, lighting conditions, face pose, gender and ethnicity [10]. Each image in the dataset contains a single face, already cropped. Considering that most of the samples in the dataset are approximately 60 × 60 pixels, we decided to discard all images smaller than 30 × 30 pixels before training, in order to not damage the regression models. The distribution of the dataset samples by age is shown in Fig. 1. This distribution shows a strong imbalance between different ages, whit under-represented ages, at the tails of the distribution, counting only 2-10 images per age. To handle this imbalance, offline and online data augmentation techniques were applied. In particular, horizontal flipping, random zooming, rotation and brightness variation were used. To adapt EfficientNet to age estimation, the network was trained by fine-tuning the weights pre-trained on ImageNet on the MIVIA Age Dataset, unfreezing the top 20 layers. The Adam optimizer was used, with a learning rate of 10^{-4}, a mini-batch size of 128, and the mean absolute error (described below) as a loss function. The dataset has been split into 80% for training and 20% for validation. To evaluate the effectiveness of the trained model, we submitted our model to the challenge organizers and they ran it on the private test set.

4.2 Metrics

Concerning face detection, we measured standard metrics commonly used in object detection tasks, i.e. precision, recall and average precision (AP). Precision is the ability of the method to detect *only* faces; in other words, it is the percentage of faces correctly detected over all detections. Recall is the ability of the method to find *all* faces; in other words, it is the percentage of faces correctly detected over all ground truth bounding boxes. Since precision-recall curves typically follow zigzag patterns, AP is also considered, which is basically the area under the precision-recall curve; it has been calculated at an intersection over union of 0.5.

Concerning age estimation, the experimented models have been evaluated using the metrics indicated by the challenge evaluation protocol: mean absolute error (MAE), regularity and a new index called *age accuracy and regularity* (AAR). The purpose of MAE is quite simple: it measures the mean absolute error of the age estimates and is defined as $MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - r_i|$, where N is the test set size, p_i is the predicted age and r_i is the ground truth age. Due to the wide age range (1–70+), it is useful to understand whether the model performance is consistent across all age groups. For this reason, a regularity score is adopted. First, ages are divided into 8 categories (1–10, 11–20, 21–30, . . .), then the MAE is calculated for each of them. Regularity is then defined as a standard deviation calculated with the rule $\sigma = \sqrt{\frac{\sum_{j}^{8}(MAE^{j} - MAE)^2}{8}}$, where MAE^j is the MAE computed on the j-th age category. The smaller the standard deviation obtained with a method, the lower its regularity. Finally, to have a summary of these metrics, AAR is used, which is defined as $AAR = \max(0, 7 - MAE) + \max(0, 3 - \sigma)$. The AAR metric varies between 0 and 10, weighing 70% the MAE contribution and 30% the regularity contribution. A perfect method, achieving $MAE = 0$ and $\sigma = 0$, will therefore get $AAR = 10$.

4.3 Results

The overall system has been tested on a PC desktop with a Ryzen 5 3600 CPU and 32GB of RAM. Using YOLOv5s, face detection on the CPU takes 0.06 seconds ($\sim$16 fps). On Google Colab, with a Tesla T4, the inference time takes 0.009 seconds ($\sim$111 fps). As for EfficientNet, the neural network is capable of processing 76 facial images per second on the CPU, and 549 images on the GPU.

As for face detection with YOLOv5s, after convergence the model achieves the results reported in Table 1. We did not compare this version of YOLO with any other or any other object detection system, as the main focus of the challenge was age estimation. The model achieved a very high precision rate, at the expense of a lower recall, which could be due to the presence of many training samples containing groups of people, sometimes with hundreds of small faces per image. However, it should be noted that this drawback is largely mitigated by the very high detection rate, which allows the model to recover a face missed in a given frame in a subsequent frame.

As for age estimation, we compared EfficientNet-B0 with a custom baseline convolutional neural network (CNN), consisting of three convolutional layers (with an ascending depth of 96, 256 and 384 filters and a descending kernel size of 7×7, 5×5 and 3×3) and two fully-connected layers on top with a dropout in between. Moreover, we made a comparison with the popular ResNet50 [14]. The preliminary results obtained on the validation set are reported in Table 2. It can be seen that both EfficientNet and ResNet outperform the baseline model. They also show comparable regularity scores. However, although the regularity

Table 1. Face detection results.

Precision	Recall	AP
94.6%	63.1%	67.6%

Table 2. Comparison of age estimation models on our validation set.

Model	MAE	Regularity	AAR
Baseline CNN	4.13	0.75	5.12
ResNet50	3.17	0.68	6.15
EfficientNet-B0	2.50	0.70	6.80

Table 3. Age estimation challenge results.

Model	MAE	Regularity	AAR
EfficientNet-B0	2.89	1.70	5.41

achieved by ResNet is slightly lower, EfficientNet shows a much better MAE. All in all, considering that EfficientNet is much more computationally efficient than ResNet, this allowed us to choose it for the participation in the challenge.

The results obtained on the competition private test set, reported in Table 3, are promising. The MAE obtained is slightly higher than that achieved in the preliminary evaluation on the validation set, indicating a small overfitting. Conversely, the regularity is much higher, indicating that the model is not stable across age groups. This could indicate that the data augmentation used to balance the dataset has not proved effective enough. To conclude our analysis, we compared the proposed method with existing deep learning-based methods proposed in the literature (see Table 4). As the challenge results are not yet publicly available at the time of writing, we have made the comparison with recent methods tested on other popular datasets, in particular FG-NET and MORPH. The comparison is less fair; however, at least we understand in what order of magnitude the error committed in the state-of-the-art is placed. Considering that state-of-the-art models do not fall below ~ 2 of MAE, that the MIVIA Age Dataset is an “in-the-wild” dataset, with noise and incorrectly labeled samples, and that the proposed method represents a very efficient solution compared to other methods, the results look pretty good.

Examples of application of the proposed system in real-time during the processing of the video stream of a simple webcam are shown in Fig. 4.3.

5 Conclusion

In this work, we have presented our system based on YOLOv5 and EfficientNet that performs face detection and age estimation from facial images in a real-time two-step process. In particular, a version of EfficientNet, fine-tuned on the new, challenging MIVIA Age Dataset, proved to be quite effective in providing an acceptable age estimation on average. As a future work, we wish to experiment

Table 4. Comparison with state-of-the-art methods based on deep learning. Only the MAE is compared, being the main metric used by previous works.

Reference	Method	Dataset	MAE
Yi et al. [29]	Multi-scale CNN	MORPH	3.63
Wang et al. [27]	CNN + dimensionality reduction	FG-NET	4.26
Niu et al. [19]	Multi-output CNN	MORPH	3.27
Rothe et al. [24]	Fine-tuned VGG16	MORPH	2.68
Othmani et al. [20]	Fine-tuned Xception	MORPH	2.01
This work	Fine-tuned EfficientNet	MIVIA Age	2.89

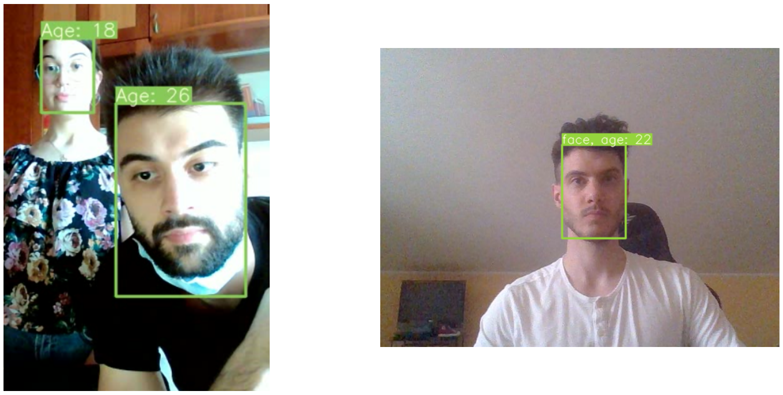

Fig. 2. Examples of applications of the proposed system in real-time with a simple webcam. The real ages are 17, 23 and 23 from left to right respectively.

with more advanced balancing and data augmentation techniques to improve system stability.

References

1. Al-Shannaq, A.S., Elrefaei, L.A.: Comprehensive analysis of the literature for age estimation from facial images. IEEE Access **7**, 93229–93249 (2019)
2. Buono, P., Castellano, G., De Carolis, B., Macchiarulo, N.: Social assistive robots in elderly care: Exploring the role of empathy. In: 1st International Workshop on Empowering People in Dealing with Internet of Things Ecosystems, EMPATHY 2020. pp. 12–19 (2020)
3. Can Malli, R., Aygun, M., Kemal Ekenel, H.: Apparent age estimation using ensemble of deep learning models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 9–16 (2016)
4. Carletti, V., Greco, A., Percannella, G., Vento, M.: Age from faces in the deep learning revolution. IEEE transactions on pattern analysis and machine intelligence **42**(9), 2113–2132 (2019)
5. Castellano, G., De Carolis, B., Macchiarulo, N., Vessio, G.: Pepper4Museum: Towards a human-like museum guide. In: Proceedings of the AVI2CH Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage (2020)

6. Fujiwara, T., Koshimizu, H.: Age and gender estimations by modeling statistical relationship among faces. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. pp. 870–876. Springer (2003)
7. Gao, F., Ai, H.: Face age classification on consumer images with gabor feature and fuzzy LDA method. In: International Conference on Biometrics. pp. 132–141. Springer (2009)
8. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on pattern analysis and machine intelligence **29**(12), 2234–2240 (2007)
9. Greco, A., Saggese, A., Vento, M.: Digital signage by real-time gender recognition from face images. In: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT. pp. 309–313. IEEE (2020)
10. Greco, A., Saggese, A., Vento, M., Vigilante, V.: Effective training of convolutional neural networks for age estimation based on knowledge distillation. Neural Computing and Applications pp. 1–16 (2021)
11. Gunay, A., Nabiyev, V.V.: Automatic age classification with LBP. In: 2008 23rd International Symposium on Computer and Information Sciences. pp. 1–4. IEEE (2008)
12. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
13. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 112–119. IEEE (2009)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Jocher, G., Stoken, A., Borovec, J., NanoCode012, Chaurasia, A., TaoXie, Changyu, L., V, A., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Hajek, J., Diaconu, L., Marc, Kwon, Y., oleg, wanghaoyang0106, Defretin, Y., Lohia, A., ml5ah, Milanko, B., Fineran, B., Khromov, D., Yiwei, D., Doug, Durgesh, Ingham, F.: ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations (Apr 2021). https://doi.org/10.5281/zenodo.4679653, https://doi.org/10.5281/zenodo.4679653
16. Kwon, Y.H., da Vitoria Lobo, N.: Age classification from facial images. Computer vision and image understanding **74**(1), 1–21 (1999)
17. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **34**(1), 621–628 (2004)
18. Li, S., Xing, J., Niu, Z., Shan, S., Yan, S.: Shape driven kernel adaptation in convolutional neural network for robust facial traits recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 222–230 (2015)
19. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4920–4928 (2016)
20. Othmani, A., Taleb, A.R., Abdelkawy, H., Hadid, A.: Age estimation from faces using deep learning: A comparative analysis. Computer Vision and Image Understanding **196**, 102961 (2020)

21. Pinter, A.T., Wisniewski, P.J., Xu, H., Rosson, M.B., Caroll, J.M.: Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In: Proceedings of the 2017 Conference on Interaction Design and Children. pp. 352–357 (2017)
22. Punyani, P., Gupta, R., Kumar, A.: Neural networks for facial age estimation: a survey on recent advances. Artificial Intelligence Review **53**(5), 3299–3347 (2020)
23. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
24. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 10–15 (2015)
25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
26. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
27. Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: 2015 IEEE Winter Conference on Applications of Computer Vision. pp. 534–541. IEEE (2015)
28. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5525–5533 (2016)
29. Yi, D., Lei, Z., Li, S.Z.: Age estimation by multi-scale convolutional network. In: Asian conference on computer vision. pp. 144–158. Springer (2014)