International Journal of Decision Science
Volume 9 • Number 1 • June 2018

A Novel Pet face Recognition Algorithm based on Improved Convolutional Neural Network

Sui Ting-Ting, Jin Yu-Ruo, Yang Yi-Fan, Miao Jun-Xiang, Zhu Jing-Run and Xia Fu-Liang-Yi
Shanghai DianJi University, China
E-mail: suisui61@163.com

ABSTRACT

Pets are becoming more and more popular, but pet feeding systems are often manual, time-consuming and labor-intensive. Therefore, this paper proposes a novel pet face detection algorithm based on improved convolutional neural network. In this algorithm, an image locating and correction process is added into the CNN model to reduce the influence of background noise and pet postures. The experimental results were validated in animals such as dogs and cats in Pascal 2012 and Google image library. The experimental results show superiority over other available approaches on recognition accuracy.

Key Words: Pet face recognition, CNN, image locating, image correction

INTRODUCTION

With the progress of society and the improvement of people's living standard, more and more people like to keep pets. However, pet feeding has brought a lot of trouble. For many pet families, the food quantity of traditional feeding machine can only be set manually (Lowe, 2010), which may cause food waste and food shortage. In this paper, a novel pet face detection algorithm based on improved convolutional neural network is proposed. By using this algorithm, the category of animals can be recognized to provide the information for later distribution of food.

At present, face recognition system has achieved many excellent results compared with pet recognition (Vinay *et al*., 2018, Dave *et al*., 2018, Wang *et al*., 2017, Luan *et al*., 2017, Kim *et al*., 2018). Among them, the most traditional face recognition method is based on geometric features (Fang *et al*., 2011, Ballihi *et al*., 2012). In this method, the face is characterized by the local shape of the eyes, nose, mouth and their geometric relationships. However, for the sake of accuracy, a lot of priori knowledge such as face structure is often needed. After that, researchers put forward a method based on templates (Karungaru *et al*., 2004, Franco *et al*., 2010). The core of it is to transform the whole image into a gray level template, forming a two-dimensional matrix, and then use the appropriate scale to match the test samples with the template. The most commonly used features are facial features. This method can achieve face recognition better, but it is easy to be affected by face shape and light conditions. For this reason, a model-based approach is proposed (Teijeiro *et al*., 2011). This method uses mathematical model to merge the information of different face instances with different facial scales and directions, which has greater flexibility for natural face deformation and illumination conditions.

Nowadays, deeplearning methods are widely used in the face recognition field Hu *et al*., 2015, Gao *et al*., 2015, Ghazi *et al*., 2016, Han *et al*., 2018, Goswami *et al*., 2018). Because deeplearning method is used as a detection to find the basic and essential parts of the objects. It can simulate a variety of human behaviors, and form a complete identification system to find common features from complex image data. Zhang *et al*. (2014) proposed a multi-task face detection method based on CNN, which improved the performance of face detection by constructing the joint learning of two auxiliary tasks: face pose estimation and face key point detection. Wang *et al*.

(2015) also used CNN to extract multi-level features for fusion and for face age estimation. Kahou *et al*. (2015) proposed a multi-modal deep network for video expression recognition, which uses DCNN to extract face feature information from video and uses DBN to extract voice information from video.

However, it is difficult to directly apply face detection to pets, because animals do not have a relatively fixed intra-class structure like human faces. At the same time, pet posture is unstable (Richter *et al*., 2010, Flexer *et al*., 2012). Sometimes they are squinting, drooping ears, curled up or lying on the ground, which are hard to extract the stable features. Especially, the image background can undoubtedly add to the difficulty of pet recognition, such as desks, blankets, sofas, kitchens, grasslands, bushes and so on. In order to simple such problems and improve the recognition results, we limit the recognition region to the common attributes between pets and we human. In general, animal faces are closest to human faces. Therefore, this paper proposes an improved CNN model, which adds the previous face locating steps and image correction steps to reduce the impact of pet recognition caused by different postures and background changes.

CONVELUTIONAL NEURAL NETWORK

Deep convolution neural network (Mayya *et al*., 2016, Zeng *et al*., 2013, Mahmoud *et al*., 2013) is the first learning algorithm to successfully train multi-layer neural network. It can effectively and independently learn image features, so it has been widely used. CNN model is based on local connection, weight sharing and sub-sampling. By optimizing the structure of the normal neural network, CNN can reduce the number of neurons and weights. In the meantime, the pooling operation makes the features have displacement, scaling and distortion invariance (Lecun *et al*., 2015). The typical deep convolution network structure contains 5 layers. The first layer is the image input layer, which is then composed of multiple convolution layers and sub-sampling layers, and the last layer is the fully connected layer.

The learning of C level

The ***C*** layer mainly uses convolution kernel to extract features, which can achieve the effect of filtering and intensification of features. In each convolution layer, convolution operation is performed between the feature maps of the previous layer and the convolution kernel. Then the feature map y_{ti} of this layer can be yield out by activation function, as shown in equation (1).

$$y_j^t = f\left(\sum_{i\in P_j} k_{i,j}^t * y_i^{t-1} + b_j^t\right) \quad (1)$$

In this formula, $f(\cdot)$ is the activation function, and the Sigmoid function is chosen as the activation function. t is used to represent the number of layers. $k_{i,j}$ is the convolution kernel, and * represents the 2D convolution operation. b_j is the bias, and P_j represents the set of selected feature maps.

The learning of S level

S layer can reduce the feature dimension of ***C*** layer by down-sampling. It will carry out the "pool average" or "pool maximum" operation on each pool with $n \times n$ size regions in ***S*** layer, then the sampling features can be obtained, as shown in equation (2).

$$y_j^t = f\left(down\left(y_i^{t-1}\right)\cdot w_j^t + b_j^t\right) \quad (2)$$

In this formula, w is the weight and $down(\cdot)$ is the down-sampling function. Moreover, the pool maximization operation is adopted. By pooling operation, the complexity of ***C*** layer is effectively reduced, and the over-fitting phenomenon is restrained. At the same time, the tolerance of features to small distortion and rotation is improved, and the performance and robustness of the algorithm are enhanced.

PROPOSED METHODOLOGY

Overall Model

Since CNN can extract more responsive features of pet images, we choose CNN structure with 7 layers (except the input layer and output layer) as the basic model. It involves both 2D convolutional layers and pooling layers. The 2D convolutional layer extracts features for jointly capturing the appearance, and is followed by a max-pooling operator to improve the robustness against local deformations and noise.

Of course, in view of the variety of pet postures and the complexity of image background, we add the pet locating and image standardization operations as the part of image pre-processing. The output of this part will be the input of CNN model. The overall structure is shown in Figure 1.

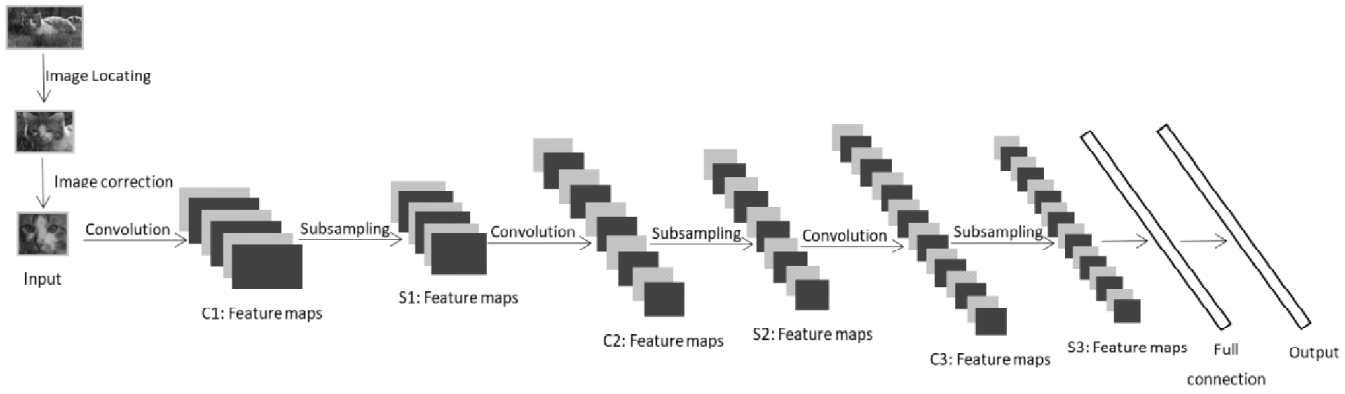


Figure 1: The structure of our method

Image locating process

Referring to the achievements of face recognition, we can find that the eyes, nose, mouth, eyebrows and other prominent facial features are often used as the basis for facial location. Therefore, the corresponding features of pet face are selected as the location basis. However, too many location features will also affect the efficiency of recognition. In order to balance the location time and locating result, we choose the eyes, nose and ears of pets as the location features.

In order to locate pet face better, we use 9 key points to locate facial features for learning. The key points are separated into two parts, which is shown in Figure 2. Each ear uses 3 key points for location. The eyes and nose use the last 3 key points. Figure 2 shows that the eyes, nose and ears depicted by the 9-point calibration method are triangular. Therefore, we use the sliding window method to get the triangle area of the ears, eyes and nose. Finally, determine the pet face according to the location relationship between the ears, eyes and nose. As shown in the rectangle box.

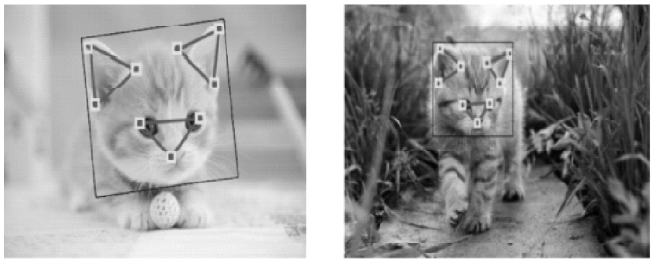


Figure 2: The cat head image is labeled by 9 points.

Image correction process

When using geometric features to recognize the images, the most important step is to properly standardize images. In order to get more precise pet face area, two eyes of pets are used as the basis for locating. Then we segment the pet images and only take the pet face area including eyes, nose and chin the positive sample.

Moreover, the posture of pets also has a certain impact on the recognition results. Therefore, we limited the pose of pets. The standard of rotation is as follows:

a In-plane rotation [-20~+20]:

$$lx = cx - 6 \times (dist/6) \quad (3)$$

$$rx = cx + 6 \times (dist/6) \quad (4)$$

$$ty = cy - 3 \times (dist/6) \quad (5)$$

$$dy = cy + 7 \times (dist/6) \quad (6)$$

b Plane to left rotation [20~60] and plane to right rotation [20~60]

$$lx = cx - 5 \times (dist/5) \quad (7)$$

$$rx = cx + 5 \times (dist/5) \quad (8)$$

$$ty = cy - 4 \times (dist/6) - dist/12 \quad (9)$$

$$dy = cy + 1 \times dist - dist/12 \quad (10)$$

In these formulas, ***lx*** represents the leftmost abscissa of the sample rectangle. ***rx*** represents the

rightmost abscissa of the sample rectangle. ***ty*** represents the top of the ordinate of the sample rectangle, and ***dy*** represents the bottom of the ordinate of the sample rectangle. So, the four vertex coordinates of the sample rectangle are (***lx***, ***ty***), (***rx***, ***ty***), (***lx***, ***dy***) and (***rx***, ***dy***). ***dist*** denotes the Euclidean distance between the centers of two eyes. (***cx, cy***) denotes the midpoint coordinates of the line between the eyes.

Besides, illumination processing, and normalization processing should be performed on the intercepted images.

The steps of training and experimental

Training stage

Step 1. According to the learning samples, the location relationship between the ear triangle, the eye-nose triangle and the pet face was calculated by the sample statistical analysis method.

Step 2. Image correction based on formula (3~10).

Step 3. Use CNN model to get the feature vector of the target.

Step 4. Train different feature vectors to get the parameters of the model.

Testing phase

Step 1. Obtain the ear triangle regions and the eye-nose triangle region from the input test images.

Step 2. Determine the area of pet face according to the records of location relations and the newly obtained triangle regions.

Step 3. Image correction based on formula (3~10).

Step 4. Use the model parameters and CNN model to identify the pet faces and output the pet categories.

RESULTS

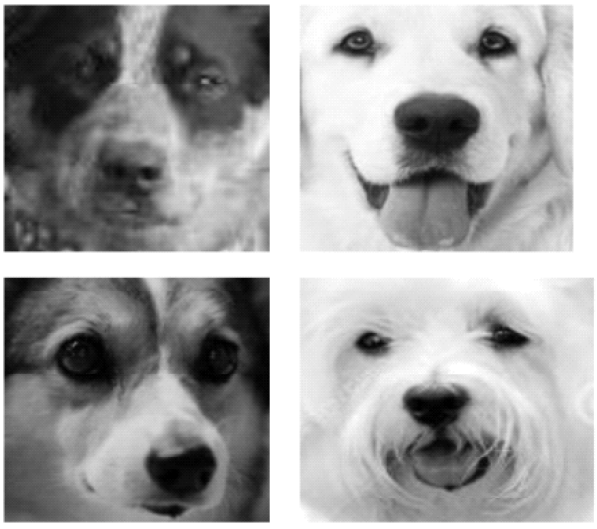
Datasets and standards

The simulation platform is configured with Core 4-core processor and a memory with 8 GB. We choose Animal images in Pascal VOC2012 as the experiment dataset, including: bird, cat, cow, dog, horse and sheep. As CNN need huge amounts of data to guarantee the recognition results, we use Google to expand the amount of image library of each class to 3000 images. The initial learning rate is set at 0.1 and linearly decreases during the iteration to find the optimal value. Meanwhile, in order to evaluate the recognition effect, a ten-fold crossover experiment is used to verify the method, and the recognition accuracy is used as the evaluation criterion, as shown in equation (11).

$$PVal_i = \frac{PT_i}{PT_i + FT_i} \tag{11}$$

In this formula, $PVal_i$ represents the accuracy of class i image recognition, PT_i represents the number of correctly identified samples, and FT_i represents the number of incorrectly identified samples.

a) Original images



b) Images after locating operation

Figure 3: The difference between original images and images after locating operation

Image locating effect

From Figure 3, we can find that original images contain much complex backgrounds. Some background information even takes up most of the image. If these images are set as input data for the training model, they will definitely affect the recognition results. However, we can find that the images after locating operation contain the main features of the pet faces. They can screen out the backgrounds which may be useless or even disturbing. These results show that the image locating process is essential for later recognition.

Image correction effect

In order to reduce the influence areas in the positioned images, we still need to reduce the scopes of the images, while taking rotation and photometric adjustment operations.

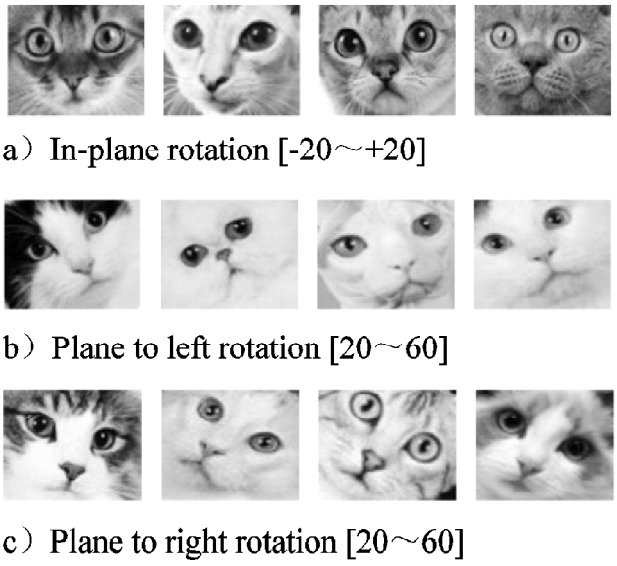


Figure 4: Image correction effect

From Figure 4, it can be found that the difference between each type of image is greatly reduced after rotation and photometric adjustment process of the positioned image, which makes the feature extraction more convenient. Therefore, the image correction process is essential for later recognition.

Image locating and correction effect

From Figure 5, it can be found that the recognition effect of CNN model with original images is obviously poor, and susceptible to the interference of external factors such as backgrounds. It shows that image locating is needed to screen out the useless and interference parts from the original images. The CNN+IPO model adds the image locating process into the CNN model, which achieves better recognition results, as shown in Figure 5.

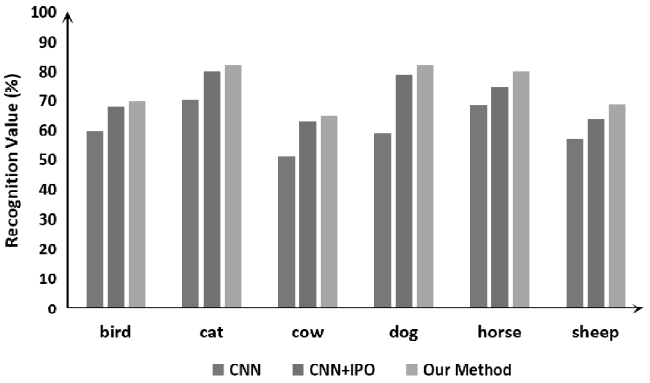


Figure 5: The recognition performance of CNN model, CNN+IPO and our method

However, geometric features are sensible to the location relationship among facial features. Therefore, our method fully considers the location relationship among the facial features. The image correction process is added into the CNN+IPO model, which improves recognition effect and enhances the robustness of our method.

Image Recognition effect

In order to verify the validity of the proposed method, our method is compared with the method in reference (Felzenszwalb *et al*., 2010, Karaoglu *et al*., 2012, Alexiou *et al*., 2014, Perronnin *et al*., 2010), as shown in Figure 6. Our method has some advantages in cat,

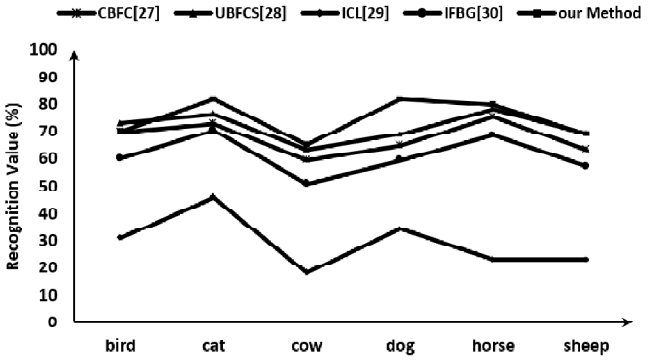


Figure 6: Recognition performance of different methods

dog horse and sheep. The reason is that these targets have notable facial features like ears, eyes and nose. Therefore, it is easier for our method to locate the face region of pets. For the methods in reference (Felzenszwalb *et al*., 2010, Karaoglu *et al*., 2012, Alexiou *et al*., 2014, Perronnin *et al*., 2010), they focus on the whole pets' image. However, pet postures in those test images are unstable. Sometimes they are squinting, drooping ears or lying on the ground, which are hard to extract the stable features. Therefore, the recognition effect is very dependent on animal posture in data set. It can be found from Figure 6 that the recognition effect of our method is superior to other methods and provides a more effective method for pet recognition.

CONCLUSION

In this paper, an improved convolutional neural network is proposed to recognition pet face. Compared with existing methods, our method has the following prominent features: 1) To imitate the process of human brain vision cognition, an image location process is added. Therefore, the recognition interference caused by the pseudo-target region are reduced. 2) The image correction process is used to reduce the ambiguity of image feature expression caused by various pet postures. In summary, our method provides a more effective way for pet recognition. The recognition results can also be used to provide a basis for later distribution of food.

ACKNOWLEDGE

This work was supported by 1) scientific research start-up fees (A1-0227-18-009-09); 2) the funding scheme for training young teachers in colleges and universitiesÿ3) science and technology innovation project for university students (A1-0224-18-012-066).

REFERENCES

1. Lowe J A. (2010). Pet rabbit feeding and nutrition. Nutrition of the Rabbit, 294-313.
2. Vinay A, Akshaykanth D L, Kamath A, *et al*. (2018). GOLD Features with Machine Learning for, Face Recognition. Social Science Electronic Publishing.
3. Dave R, Vyas A, Mojidra S, *et al*. (2018). Face Recognition Techniques: A Survey.
4. Wang W, Wang R, Huang Z, *et al*. (2017). Discriminant Analysis on Riemannian Manifold of Gaussian Distributions for Face Recognition with Image Sets. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, (99): 1-1.
5. Luan T, Yin X, Liu X. (2017). Disentangled Representation Learning GAN for Pose-Invariant Face Recognition, Computer Vision and Pattern Recognition. IEEE, 1283-1292.
6. Kim K G, Yang Z, Masi I, *et al*. (2018). Face and Body Association for Video-Based Face Recognition, Applications of Computer Vision. IEEE, 39-48.
7. Fang Y U, Yue D L, Wang S H. (2011). Face Recognition Algorithm Based on Improved Geometric Features. Computer Simulation, 28(4):291-294.
8. Ballihi L, Amor B B, Daoudi M, *et al*. (2012). Boosting 3D Geometric Features for Efficient Face Recognition and Gender Classification. IEEE Transactions on Information Forensics & Security, 7(6): 1766 -1779.
9. Karungaru S, Fukumi M, Akamatsu N. (2004). Face recognition using genetic algorithm based template matching, IEEE International Symposium on Communications and Information Technology. IEEE, 2:1252- 1257.
10. Franco A, Maio D, Maltoni D. (2010). Incremental template updating for face recognition in home environments. Pattern Recognition, 43(8):2891-2903.
11. Teijeiro Mosquera L, Alba-Castro J L. (2011). Performance of active appearance model-based pose-robust face recognition. Iet Computer Vision, 5(6):348-357.
12. Huang Y S, Chen S Y. (2015). A geometrical-model-based face recognition, IEEE International Conference on Image Processing. IEEE, 3106-3110.
13. Hu G, Yang Y, Yi D, *et al*. (2015). When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition 384-392.
14. Gao S, Zhang Y, Jia K, *et al*. (2015). Single Sample Face Recognition via Learning Deep Supervised Autoencoders. IEEE Transactions on Information Forensics & Security, 10(10):2108-2118.

15. Ghazi M M, Ekenel H K. (2016). A Comprehensive Analysis of Deep Learning Based Representation for Face Recognition, Computer Vision and Pattern Recognition Workshops. IEEE, 102-109.
16. Han X, Du Q. (2018). Research on face recognition based on deep learning, Sixth International Conference on Digital Information, Networking, and Wireless Communications. IEEE.
17. Goswami G, Ratha N, Agarwal A, *et al*. (2018). Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks.
18. Zhang C, Zhang Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks, Applications of Computer Vision. IEEE, 1036-1041.
19. Wang X, Guo R, Kambhamettu C. (2015). Deeply Learned Feature for Age Estimation, IEEE Winter Conference on Applications of Computer Vision. IEEE Computer Society, 534-541.
20. Kahou S E, Bouthillier X, Lamblin P, *et al*. (2015). EmoNets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces, 10(2):1-13.
21. Richter T, Bergmann R, Pietzsch J, *et al*. (2010). Effects of posture on regional pulmonary blood flow in rats as measured by PET. Journal of Applied Physiology, 108 (2):422.
22. Flexer H, Heimlich E. (2012). Device, system and method for monitoring animal posture pattern: US, US 8111166 B2.
23. Mayya V, Pai R M, Pai M M M. (2016). Automatic Facial Expression Recognition Using DCNN. Procedia Computer Science, 93: 453-461.
24. Zeng H, Edwards M D, Liu G, *et al*. (2016). Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics, 32(12): i121-i127.
25. Mahmoud K M. (2013). Handwritten Digits Recognition using Deep Convolutional Neural Network: An Experimental Study using EBlearn. Computer Science.
26. Lecun Y, Bengio Y, Hinton G. (2015). Deep learning. Nature, 521(7553):436.
27. Felzenszwalb P F, Girshick R B, Mcallester D, *et al*. (2010). Object Detection with Discriminatively Trained Part-Based Models. IEEE Transactions on Pattern Analysis & Machine Intelligence, 32(9): 1627-1645.
28. Karaoglu S, Gemert J C V, Gevers T. (2012). Object reading: text recognition for object recognition, International Conference on Computer Vision. Springer Verlag, 456- 465.
29. Alexiou I, Bharath A A. (2014). Spatio chromatic Opponent Features, European Conference on Computer Vision. Springer International Publishing, 81-95.
30. Perronnin F, Sánchez J, Mensink T. (2010). Improving the Fisher Kernel for Large-Scale Image Classification. Eccv, 115 (7): 143-156.