Hybrid Learning Framework for Large-Scale Web Image Annotation and Localization

Yong Li[1], Jing Liu[1], Yuhang Wang[1], Bingyuan Liu[1], Jun Fu[1], Yunze Gao[1], Hui Wu[2], Hang Song[1], Peng Ying[1], and Hanqing Lu[1]

[1]IVA Group, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] Institute of Software, Chinese Academy of Sciences
{yong.li,jliu,yuhang.wang,byliu,peng.ying,luhq}@nlpr.ia.ac.cn
{fujun2015,gaoyunze2015}@ia.ac.cn
wuhui13@iscas.ac.cn,hangsongv@gmail.com
http://www.nlpr.ia.ac.cn/iva
http://www.foreverlee.net

Abstract. In this paper, we describe the details of our participation in the ImageCLEF 2015 Scalable Image Annotation task. The task is to annotate and localize different concepts depicted in images. We propose a hybrid learning framework to solve the scalable annotation task, in which the supervised methods given limited annotated images and the search-based solutions on the whole dataset are explored jointly. We adopt a two-stage solution to first annotate images with possible concepts and then localize the concepts in the images. For the first stage, we adopt the classification model to get the class-predictions of each image. To overcome the overfitting problem of the trained classifier with limited labelled data, we use a search-based approach to annotate an image by mining the textual information of its similar neighbors, which are similar on both visual appearance and semantics. We combine the results of classification and the search-based solution to obtain the annotations of each image. For the second stage, we train a concept localization model based on the architecture of Fast R-CNN, and output the top-k predicted regions for each concept obtained in the first stage. Meanwhile, localization by search is adopted, which works well for the concepts without obvious objects. The final result is achieved by combing the two kinds of localization results. The submitted runs of our team achieved the second place among the different teams. This shows the outperformance of the proposed hybrid two-stage learning framework for the scalable annotation task.

Keywords: Hybrid Learning, SVM, Fast R-CNN, Annotation, Concept Localization

1 Introduction

With the advance of digital cameras and high quality mobile devices as well as the Internet technologies, there are increasingly huge number of images available on the web. This necessitates scalable image annotation techniques to effectively organize and retrieval the large scale dataset. Although some possibly related textual information to images is presented on their associated web pages, the relationship between the surrounding text and images varies greatly, with much of the text being redundant and unrelated. Therefore, how to best explore the weak supervision from textual information is a challenging problem for the task of scalable image annotation.

The goal of scalable image annotation task in ImageCLEF 2015 is to describe visual content of images with concepts, and to localize the concepts in the images [7, 17]. The task provides a dataset of 500,000 web images with the textual information extracted from web pages, in which 1979 items with ground truth localized concept labels form the development set. The overall performance will be evaluated to annotate and localize concepts on the full 500,000 images. The large scale test data and the new task of concept localization are the main differences from the previous ImageCLEF challenges. Unlike the other popular challenges like ILSVRC [14] and Pascal [5], such task has few fully labelled training data but a large-amount of raw web resources used for model learning.

For the participation of the scalable image annotation task, we adopt a two-stage hybrid learning framework to fully use the limited labelled data and the large scale web resource. In the first stage, we train a SVM classifier for each concept in a one-vs-rest manner. To avoid the overfitting problem brought by the small scale training data, we adopt another unsupervised solution as a complement to enhance the scalability of our work. We attempt to annotate an image by search on the whole 500,000 dataset, in which the visual and semantical similarities are jointly estimated with deep visual features [10] and deep textual features (i.e., Word2Vec [12]), and the WordNet is used to mine the relevant concepts from the textual information of those similar images. After the concept annotation stage, we obtain a set of concepts relevant to each image. We will localize the concepts through the second stage, in which the latest deep model, Fast R-CNN [8] is adopted to predict the possible locations of the concepts obtained in the first stage. Although the deep model can directly predict and localize the concepts depicted in each images, the performance is unstable possibly due to the too small number of training data with ground truth localized concept labels, which can be demonstrated from the experimental results. Thus, we use the top-K predicted regions to each concept obtained in the first stage as outputs. Besides, we adopt a search-based approach to localize the scene-related concepts (e.g., “sea”, “beach” and “river”). Specifically, the location of each predicted concept for an image is decided by the spatial layout of its visually similar images in training dataset. The experimental results show that the hybrid two-stage learning framework contributes to the improvement of image annotation and localization. Furthermore, there are a few concepts related with the concept “face” (e.g., “head”, “eye”, “nose”, “mouth” and “beard”). Since face detection and facial point detection have been actively studied over the past years and achieve satisfactory performance [19, 4, 15], we employ face detection and facial point detection to localize face related concepts exactly.

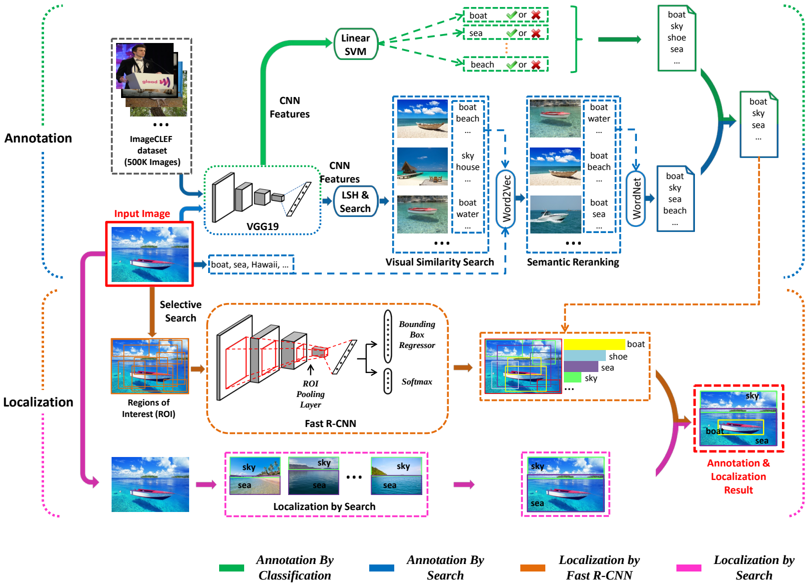

Fig. 1. Flowchart of the proposed method. The top part shows the first stage of image annotation. For a given query image, deep visual feature is extracted with the VGG 19 network. Then the SVM classifier is adopted to make prediction for the annotation. Furthermore, visual based retrieval is performed and the result is reranked with the surrounding text. The annotations to the testing image will be mined from the textual descriptions of the above obtained similar image set with the WordNet. The lower part shows the second stage of concept localization. Two methods are employed during localization. The Fast R-CNN is adopted, which works well for the concepts with obvious object. Meanwhile, localization by search method is adopted, which works well for the scene related concepts (e.g., “sea” and “beach”). Finally, localization results from the two methods are combined.

The remainder of this working note is structured as follows. Section 2 presents the details about data preparation for model training. In Section 3, we elaborate the details about how we obtain the results of image annotation. Section 4 introduces the annotation localization approach. In section 5, we discuss the results of experiments and the parameter settings. Finally, we conclude our participation in ImageCLEF 2015 in section 6.

2 Data Preparation

In this year, hand labeled data is allowed in the image annotation and localization task. We prefer multiple online resources to perform such task, including the ImageNet database [14], the Sun database [18], the WordNet [13] and the online image sharing website Flicker [1]. To perform image annotation by classification, we attempt to collect the training images from the well labeled dataset ImageNet and Sun dataset. There are 175 concepts concurrent in the ImageNet dataset and the ImageCLEF task simultaneously. Meanwhile, there are 217 concepts concurrent in the Sun dataset and the ImageCLEF task. For the concepts not in the ImageNet and Sun database, images are crawled from the online image sharing website Flicker and filtered by humans with 50 images left for each concept. In our work, the visual features of images are represented with deep features [9], we employ the VGG19 model [10] pretrained on the ImageNet dataset (1000 classes) and average the output of its *relu6* layer for 10-view image patches (4 corners and 1 center patches of an image as well as their mirrors) as our visual feature.

There are 1979 images has been released to test the proposed method. The frequency of different concept is unbalanced and there are 17 concepts do not occur in the development dataset. Then we have collected some images for such concepts to make the development set more applicable to set hyper-parameters and validate the proposed method.

3 Concept Annotation

3.1 Annotation By Classification

Image annotation by classification is to train a multi-class classifier or one-vs-rest classifier corresponding to different concepts. Such solution is simple, and usually can achieve satisfactory performance given abundant training data. Towards this problem, we choose a linear Support Vector Machine (SVM) [6] to train a one-vs-rest classifier for each concept. Due to images usually being labelled with multiple concepts in training data, the negative samples for a given concept classifier are selected as the ones whose all labels do not include the concept. For a testing image, we select the most confident concepts by thresholding the classification confidences to obtain the annotations of each image.

3.2 Annotation By Search

The search-based approach for image annotation works on the assumption that visual similar images should reflect similar semantical concepts, and most textual information of web images is relevant to their visual content. Thus, the search-based annotation process can be divided into two phases: one is the search for similar images, and the other is relevant concept selection from the textual information of those similar images.

First, given a testing image with textual information, we will search its similar neighbors on the whole 500,000 dataset. As mentioned in section 2, images are represented with 4096-dimensional deep features. To speed up similar image

[1] https://www.flickr.com/

retrieval for the large scale image database, we adopt a hash encoding algorithm. Specially, we map the deep features to 32768-dimensional binary hash codes leveraging the random projection algorithm proposed in [2], and employ hamming distance to rank the images in the dataset.

To further improve the results of visual similarity search, we explore the textual information of the given image, and perform the semantic similarity search on the top-N_A visually similar images to rerank the similar image set. Here, we use a publicly available tool of Word2Vec [12] to compute vector representations of textual information of images, which are provided with the scofeat descriptors. With the word vector representations, the cosine distance is used to rerank images in order to obtain a set of visually and semantically similar images.

Next, the annotations to the testing image will be mined from the textual descriptions of the above obtained similar image set. For the annotation mining, we employ a WordNet-based approach, which is similar to the solution in [3]. The major difference is that we mine the concepts from a set of visually and semantically similar images, while they considered only the visual similarities among images. A candidate concept graph is built with the help of WordNet, and the top-N_W concepts with higher number of links are selected as the final image description.

We combine the results of the above classification-based solution and the search-based solution with different thresholding settings, while their different performances will be discussed in the experimental session.

Concept extension is adopted to deal with the strong correlation among concepts to make the annotation result more sufficient. For the given 251 concepts, some concepts have strong correlation like “eyes” and “nose”, which usually occur together. Hierarchy relation may exist like concepts “apple” and “fruit”. When the child node concept “apple” occurs, the parent node concept “fruit” must occur. These relations can be achieved by exploring the WordNet concept hierarchy and the provided ImageCLEF concept set with general level categories.

4 Annotation Localization

4.1 Localization by Fast RCNN

To localize the objects in the given images, we follow the Fast R-CNN framework (FRCN) proposed in [8], which provides classification result and regressed location simultaneously for each candidate object proposal. The FRCN approach is conducted on a number of regions of interest (RoIs), which are sampled from the object proposals of the images. To reduce repetitive computation of the overlapped regions, the last pooling layer of the FRCN network is replaced with a RoI pooling layer compared with traditional CNN network, which observably speeds up the training and testing process. Furthermore, the network uses two sibling loss terms as supervision to learn the classification and localization information collaboratively, which are proved to be helpful to improving the performance and make the approach a one-stage detection framework.

Each RoI corresponds to a region in the feature map provided by the last convolutional layer. The RoI pooling layer carries out max pooling on each of the corresponding regions and pools them into fixed-size feature maps. The scale of pooling mask in RoI pooling layer is auto-adjusted according to the spatial size of the input feature map regions, to make the outputs all have the same size. Therefore, the feature map of each RoI can match the following fully connected layers seamlessly after RoI pooling and contribute to the network as an independent instance.

As for the supervision, the FRCN network employs two sibling output layers to predict classification probability and bounding box regression offsets respectively for each RoI on each category. The first output layer is a softmax layer which outputs a probability distribution over all categories. And we use the standard cross-entropy loss function to constrain it as follows,

$$L_{cls} = -log(\hat{p}_{k^*}) \tag{1}$$

where k^* is the groundtruth label and $\hat{p}_{k^*}$ is the predicted probability on this class, assuming that there are totally $K+1$ categories including K object classes and a background class. The second output layer is a regression layer which predicts the bounding box regression offsets for each category as $t^k = (t_x^k, t_y^k, t_h^k, t_w^k)$, where k is the index of the K object classes. Assuming that the groundtruth offsets for the class k^* is $t^* = (t_x^*, t_y^*, t_h^*, t_w^*)$, the regression loss function is formulated as follows,

$$L_{loc} = \sum_{i \in \{x,y,h,w\}} smooth_{L_1}(t_i^{k^*}, t_i^*) \tag{2}$$

where

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3}$$

Thus, the loss function for the whole network can be formulated as follows,

$$L = L_{cls} + \lambda[k^* \geq 1]L_{loc} \tag{4}$$

where λ is a weighting parameter to balance the two loss terms. And $[k^* \geq 1]$ is an indicator function with the convention that the background class is labeled as $k^* = 0$ and the object classes as $k^* = 1, 2, \cdots, K$, which means that the localization regression loss term is ignored for the background RoIs.

In practice, we first extract object proposals with the selective search method [16] and then sample RoIs from them for training. We take 25% of the RoIs from the object proposals that overlap certain groundtruth bounding boxes with more than 0.5 IoU (intersection over union) and label them the same with the groundtruth bounding boxes. The rest RoIs are sampled from the object proposals with a maximum IoU between 0.1 and 0.5 with the groundtruth bounding boxes and are labeled as background, as instructed in [8]. While during testing, we input all the object proposals into the network and predict labels and regression offsets for all of them. A preliminary screening is also implemented in this

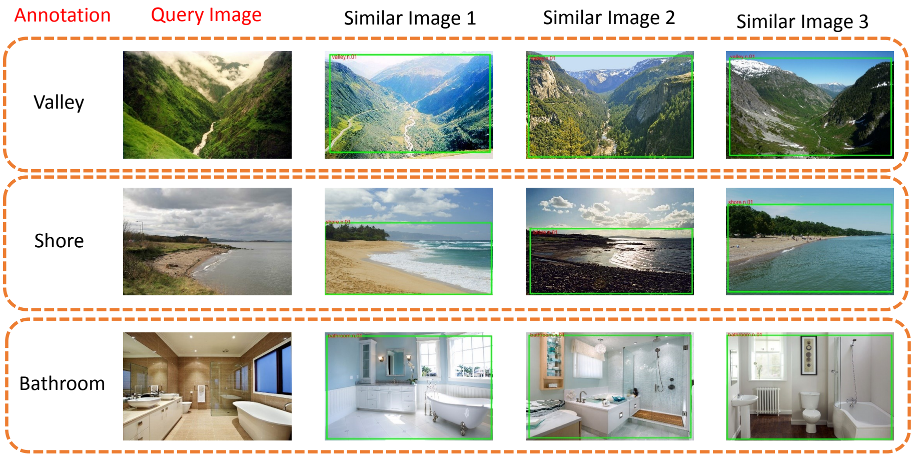

Fig. 2. Examples about search-based localization. Annotation of the query image is achieved by the method introduced in Section 3. Similar images is achieved by visual similarity. Localization of a given concept for the query image can be achieved by transferring the bounding box in the similar images.

step with non-maximum suppression to take out some object proposals with too low classification probabilities.

4.2 Localization by Search

We give a special consideration to the scene related concepts for the annotation localization (e.g., "beach", "sea", "river" and "valley"). If an image is predicted as a scenery concept, we first find its top-N_L visually similar neighbors with the same concept in the localization training data, and use their merged bounding box as the location of the scenery concept. Figure 2 shows some examples about search-based location results. The final localization results will be composed of the predicted results of the Fast R-CNN model and search-based localization results.

4.3 Localization of Face Related Concepts

There are a few concepts related with the concept "face" (e.g., "head", "eye", "nose", "mouth" and "beard"). To localize face related concepts exactly, we employ face detection algorithm with aggregate channel features [19, 4]. Facial point detection is performed to locate key points in the face [15]. Localization of the concepts "eye", "nose", "mouth" and "beard" is got based on the key points. Concepts of "ear" and "neck" are located based on the relative location of face with experience. Besides, linear classifiers [6] are trained with the SIFT [11] features extracted on the facial points to determine the concepts of "man", "woman", "male child" and "female child".

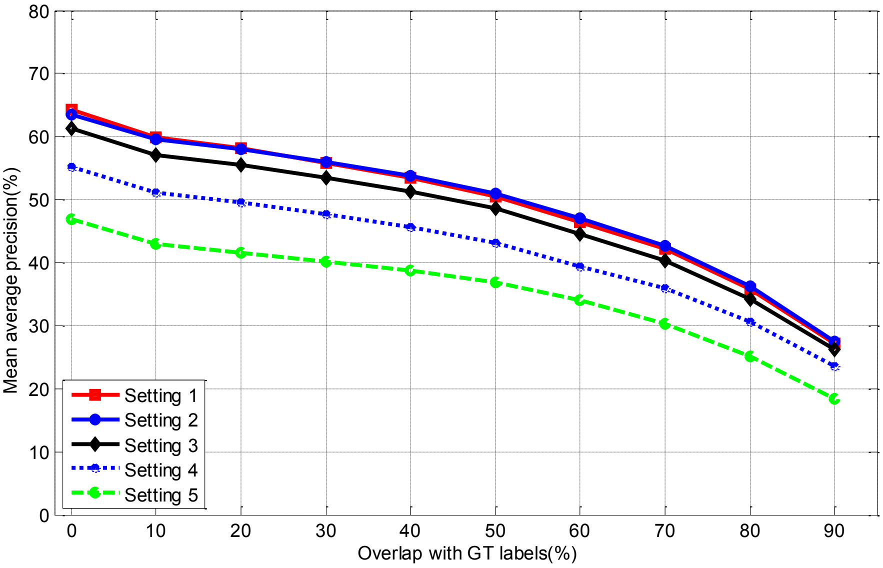

Fig. 3. Mean average precision with different recall rates.

5 Experiments

We have submitted 8 runs with 5 different settings of combinations with the above model modules, including Annotation By Classification (ABC), Annotation By Search (ABS), localization by Fast R-CNN (FRCN), Localization By Search (LBS) and Concept Extension (CE). Some runs are of the same setting with different parameter values.

Table 1. Results for the submitted runs with different settings

Method	ABC	CE	ABS	LBS	FRCN	SVM_Threshold	Overlap 0.5	Overlap 0
Setting_1	yes	no	yes	yes	yes	0.5	0.510	0.642
Setting_2	yes	yes	yes	yes	yes	0.4	0.510	0.635
Setting_3	yes	yes	no	yes	yes	0.4	0.486	0.613
Setting_4	yes	no	no	yes	yes	0.4	0.432	0.552
Setting_5	no	no	no	no	yes	0.4	0.368	0.469

The experimental results of our method with different settings are presented in Table 1. The last two columns show the mean average precision with different overlap percentage with ground truth labels. Detailed results with increasing overlap percentage is shown in Fig. 3. We can find the setting_1 with the modules ABC, ABS, FRCN and LBS achieves the best result, and setting_2 extending

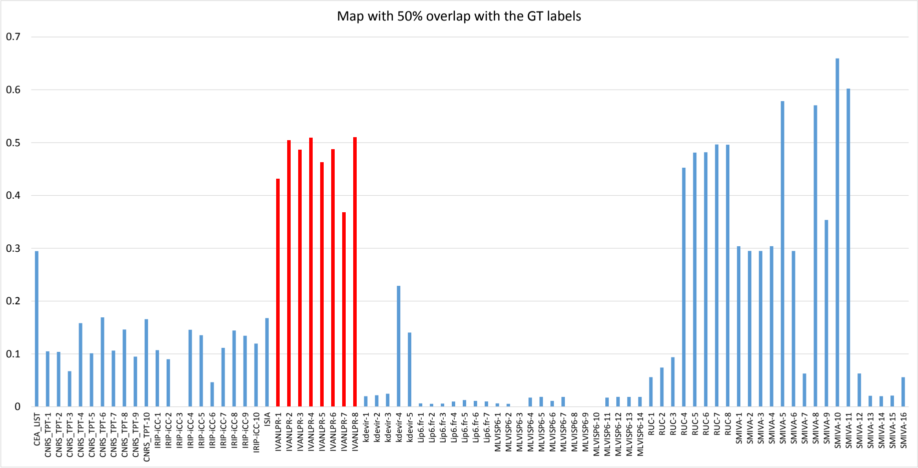

Fig. 4. Submission results of different teams with 50% overlap with the ground truth labels. Results of our method with different runs are colored red.

setting_1 with CE achieves comparable results. By comparing the result of setting_2 and setting_3, we can find the effectiveness by taking annotation by search into consideration. Furthermore, we have validated the effect of FRCN, and we find that result of localization by detection only is unsatisfactory. This is mainly due to two reasons, one is that the training data is very limited for each concept, the other is that the content in the web images do not have obvious objectness [1] for some concepts. The proposed hybrid learning framework with two stage process is more suitable to deal with such task. Comparisons of our runs (denoted IVANLPR-*) and other participants' runs are illustrated in figure 4. The submitted runs of our team achieved the second place among the different teams, which shows the outperformance of the proposed hybrid two-stage learning framework for the scalable annotation and localization task.

6 Conclusion

In this paper, we described the participation of IVANLPR team at ImageCLEF 2015 Scalable Concept Image Annotation task. We proposed a hybrid learning framework to solve the scalable annotation task. We adopt a two-stage solution to first annotate images with possible concepts and then localize the concepts in the images. For the first stage, both supervised method and unsupervised method are adopted to make full use of the available hand-labeled data and surrounding text in the webpage. For the second stage, Fast R-CNN and search-based method are adopted to locate the annotation concepts. Extensive experiments demonstrate the outperformance of the proposed hybrid two-stage learning framework for the scalable annotation task.

7 Acknowledgments.

This work was supported by 973 Program (2012CB316304) and National Natural Science Foundation of China (61332016, 61272329 and 61472422).

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. pp. 73–80 (2010)
2. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: Applications to image and text data. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 245–250. KDD, ACM (2001)
3. Budíková, P., Botorek, J., Batko, M., Zezula, P.: DISA at imageclef 2014: The search-based solution for scalable image annotation. In: Working Notes for CLEF 2014 Conference. pp. 360–371 (2014)
4. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(8), 1532–1545 (2014)
5. Everingham, M., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
6. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874 (Jun 2008)
7. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
8. Girshick, R.: Fast r-cnn. arXiv preprint arXiv:1504.08083 (2015)
9. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
10. K, S., A., Z.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
13. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM 38(11), 39–41 (1995)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (2015)
15. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR. pp. 3476–3483. IEEE (2013)
16. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104(2), 154–171 (2013)

17. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science, Springer International Publishing (2015)
18. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR. pp. 3485–3492 (2010)
19. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Aggregate channel features for multi-view face detection. In: IJCB. pp. 1–8. IEEE (2014)