Sensitive Information Detection on Cyber-Space

Mingbao Lin, Xianming Lin(✉), Yunhang Shen, and Rongrong Ji

Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Information Science and Engineering, Xiamen University,
Xiamen 361005, China
linxm@xmu.edu.cn

Abstract. The fast development of big data brings not only abundant information to extensive Internet users, but also new problems and challenges to cyber-security. Under the cover of Internet big data, many lawbreakers disseminate violence and religious extremism through the Internet, resulting in network space pollution and having a harmful effect on social stability. In this paper, we propose two algorithms, i.e., iterative based semi-supervised deep learning model and humming melody based search model, to detect abnormal visual and audio objects respectively. Experiments on different datasets also show the effectiveness of our algorithms.

Keywords: Object detect · Query by humming · Deep learning
Cyber-space security · Internet big data

1 Introduction

With the advent of big data era [5,12], the exploding information and data become increasingly aggravating. According to the statistical data by IDC, in the near future, there will be about 18EBs of storage capacity in China. The joint-report by IDC and EMC points that there will be 40000EBs globally in around 2020.

Such enormous Internet data brings not only abundant information to extensive Internet users, but also new problems and challenges to Cyber-security. Under the cover of Internet big data, many lawbreakers disseminate violence and religious extremism through the Internet. Such videos or audios are usually implanted in seemingly common data, under which it's much complicated to figure out whether it is a normal case or not. Recent years, many videos and audios in referring to violence and extreme religious beliefs have been uploaded to the Internet. These illegal data contributes a lot to the propaganda of violent events and extreme religious thoughts. How to find these illegal hidden videos or audios over mass data and get rid of them to manipulate the healthy development of Cyber-space has become a core problem to be solved immediately.

There are two types of sensitive data to be detected on the Cyber-space: one is visual objects detection, the other is audio contents detection.

Y. Zhao et al. (Eds.): ICIG 2017, Part III, LNCS 10668, pp. 496–504, 2017.
https://doi.org/10.1007/978-3-319-71598-8_44

Object detection [2,26] has been a hot topic in the field of computer vision and image processing. A lot of works about specific target detection have been done at home and abroad, e.g., pedestrian detection [6,8,18,20], vehicle detection [4,19], face detection [1,3,25], etc. By analyzing their work, we can find that early works focused on artificial definition based visual features detection. It is difficult to gain the semantic features because artificial definition based visual features have highly to do with low-level visual features. For example, Dalal and Triggs [6] raised gray gradient histogram features, which are applied to pedestrian detection. Ahonen et al. [1] presented LBP features, which are used to detect human faces. Due to the lack of interpretation of image semantics, these methods have a disappointing generalization. Recently, deep neural network has been widely applied to the domain of object detection. Not only can it learn feature descriptors automatically from object images, but also it can give a full description from low-level visual features to high-level semantics. Hence, deep learning has become popular in object detection and achieved a series of success, e.g., Tian et al. [4] transferred datasets of scene segmentation to pedestrian detection through combining the deep learning with transfer learning and gain a good achievement. Chen et al. [4] parallelized the deep convolutional network which has been applied to vehicle detection on satellite images. In [1], a deep convolutional network was proposed to detect human face with 2.9% recall improvement on FDDB datasets.

Audio retrieval has become a main direction of multi-media retrieval since the 1990s [9,14]. Based on the used data features, existing techniques are simply divided into three major categories: physical waveform retrieval [14,15], spectrum retrieval [10] and melody feature retrieval [9,11,17,23,24]. Physical waveform retrieval is time domain signal based. In [15], a prototype of audio retrieval system is designed through splitting audio data into frames with 13 physical waveform related information extracted as a feature vector and Mahalanobis distance used as a similarity metric. Spectrum retrieval is frequency domain signal based. Foote [7] extracted audio data's MFCC features and then got histogram features, which were applied to audio retrieval. In [10], a feature descriptor method based on global block spectrum has been proposed, which can present the whole spectrum information but lack anti-noise capacities. Melody feature retrieval is based on voice frequency. In 1995, Ghias et al. [9] first suggested humming melody clips to be used as music retrieval, setting a foundation of humming retrieval. McNab et al. [17] extended Ghias' idea of pitch contour and proposed to find out the continuous pitch to split notes with the help of related core technologies, like approximate melody matching or pitch tracking. Roger Jang and Gao [11], Wu et al. [24], Wang et al. [23] contributed a lot to voice frequency based melody feature retrieval successively.

In a word, with the prosperities of Internet big data, Cyber-space security are facing an increasingly serious challenge. Here are the organizations of this paper. The different bricks -sensitive visual object detection and sensitive audio information detection- are presented in Sects. 2 and 3, with proposed methods and experiments included. Conclusion are described in Sect. 4.

2 Iterative Semi-supervised Deep Learning Based Sensitive Visual Object Detection

Usually, sensitive visual information contains some particular illegal things, e.g., designated icons. Hence, to some extent, visual detection can be transformed into specific object detection.

One big obstacle of specific object detection is to grab labeled data, which is kinda a waste of human resources. What's more, human-labeled data contains noise, affecting the performance. In real life, usually we can only get data with few labeled and most unlabeled. To solve the lack of labeled data, we proposed an iterative semi-supervised deep learning based sensitive visual object detection. This algorithm can make full use of the supervised information and will focus on more and more specific objects and reinforce them as iterations.

2.1 Iterative Semi-supervised Deep Neural Network Model

Given a set of N labeled vectors

$$D = \{(x_1, y_1), ..., (x_i, y_i), ..., (x_N, y_N)\}, \tag{1}$$

among which, x_i is the ith data and y_i is its corresponding label, the learning process adjusts the set D each iteration, after which, new set is applied to update the neural network model.

First, extract M image blocks with sliding window from each training data in D. A total of $N \times M$ blocks are gained, denoted as R

$$R = \{r_{11}, ...r_{ij}, ...r_{NM}\} \tag{2}$$

Here, r_{ij} denotes the jth block from ith training data in D.

Then, classify blocks R in the neural network model learned by D and we get a new set named P, with each element a triplet

$$P = \{(r_{11}, t_{11}, s_{11}), ..., (r_{ij}, t_{ij}, s_{ij}), ..., (r_{NM}, t_{NM}, s_{NM})\} \tag{3}$$

Here, r_{ij} stands for the element in R, namely, the jth block from ith training data. And t_{ij}, s_{ij} are its corresponding class and score, resulting from the neural network learned by set D. s_{ij} is a confidence coefficient of r_{ij} belonging to class t_{ij}. Based on this, we can construct a new set D', which can be used to update the neural network model.

$$D' = \{(r_{ij}, t_{ij}) | (r_{ij}, t_{ij}, s_{ij}) \in P, t_{ij} = y_i, s_{ij} > \tau\} \tag{4}$$

This shows that the new set consists of the block that its predicted class agree with the label of its original training data and its predicted confidence coefficient exceeds a particular threshold τ.

We show a single version of our iterative model in Fig. 1 and the full algorithm is described in Algorithm 1.

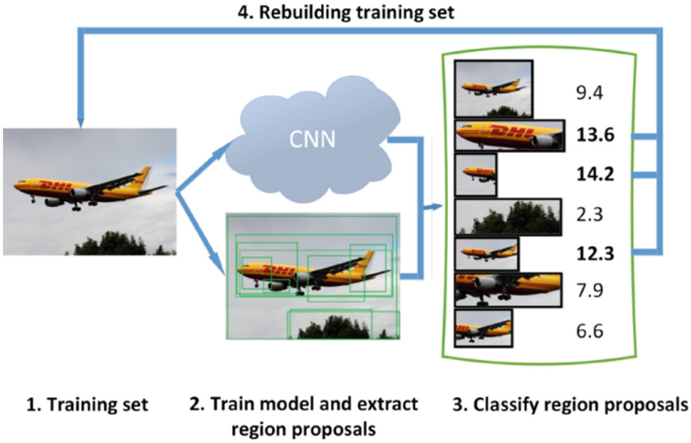

Fig. 1. An example of proposed iterative model. First, sensitive training set is collected. Then, apply this training set to training a neural network model and extract region proposals. Third, classify extracted region proposals with trained model. Lastly, rebuild the training set.

Algorithm 1. Iterative semi-supervised deep learning based Sensitive visual object detection

---

Input: pre-trained deep learning model M^0 and initial dataset D^0
Output: reinforced deep learning model
Step1: initialize No. of iterations $i \leftarrow 1$
Step2: consist of the following sub-steps
 Step2.1: $i \leftarrow i + 1$
 Step2.2: tune model M^{i-1} with dataset D^{i-1} and get a new updated model M^i
 Step2.3: according to (2), gain image blocks set R^i
 Step2.4: classify R^i with model M^i, and get P^i according to (3)
 Step2.5: get D^i based on (4)
Step3: if iteration terminates, turn to Step4, else Step2
Step4: Output latest model M^i

2.2 Experiment and Analysis

To verify the effectiveness of proposed algorithm, we compare on Flickr-32 LOGO dataset our algorithm with RCNN. This dataset contains 32 different LOGO and is split into three groups: training set, validation set and test set. Training set consists of 320 images with 10 per class. Validation set and test set consist 960 images with 30 per class, respectively. Also, we use ILSVRC2012 to pre-train CNN neural network model M^0. Selective Search Algorithm [21] is used as region proposals. For the consideration of fairness, we remove the last softmax layer and add a linear SVM. One thing should be noticed that the proposed method is kinda like RCNN. RCNN belongs to supervised algorithm, which needs the position label of LOGO, but the proposed method doesn't need.

All the experiments were complemented with python and conducted on a Dell workstation with 2 Intel E5 processors, 64G memories, 4G Navidia Quadro GPU and 8 T hard disk.

Figure 2 shows how our proposed algorithm updates the dataset. As we can see, the logo object becomes a focus as iterations with a stronger confidence coefficient.

Fig. 2. An example of proposed iterative based algorithm. As iteration goes (from (a) to (d), from (e) to (h)), object becomes clear.

We conduct experiments on Flickr dataset, and the results are compared with the art-of-state RCNN algorithm. We use mAP as an evaluation criterion. The results are shown on Table 1.

The first shows the evaluation of accuracy of R-CNN and second shows ours. In third line and fourth line, position regression are added to RCNN and proposed algorithm, denoted as R-CNN-BB and OUR-BB respectively. We should take care that the CNN network in RCNN uses 200 thousand training images for fine tune. But for our model, only 320 images are used for first fine ture and in the 12nd iteration, we acquire up to 4 thousand images. What's more, as we can see from the table, our proposed method significantly improves over RCNN, with 0.14% improvement comparing R-CNN-BB with OUR-BB.

Compared with RCNN, our proposed method - Iterative semi-supervised deep neural network model shows advances. Three general advantages are summarized as following:

First, our method can find the most stable and important inner-class features. If an image is discrete point, only a few training data can be derivated.

Second, our method has low demand on training data that there is no need to know the position of logo in the image. The training data in the next round is complemented by the confidence coefficient, while RCNN model needs strong supervised information, where positive data is defined by value of IoU (above 0.5).

Table 1. Experiment results on different logo classes comparing proposed method with RCNN

Class	Starbucks	Heineken	Tsingtao	Guiness	Corona	Adidas	Google	Pepsi	Apple	DHL	HP
R-CNN	**99.51**	**75.79**	80.58	77.72	86.83	51.30	61.28	57.19	75.77	36.94	48.33
OUR	**99.51**	73.57	81.17	73.39	88.90	**56.79**	66.40	58.80	67.89	37.95	52.78
R-CNN-BB	**99.51**	74.90	83.45	79.11	90.91	53.38	68.39	61.08	**76.12**	45.00	47.90
OUR-BB	99.28	71.03	**85.00**	**81.86**	**91.63**	53.47	**77.14**	**68.48**	73.22	**45.95**	**56.02**
Class	Rittersport	Carlsberg	Paulaner	Fosters	Nvidia	Singha	Fedex	Becks	Aldi	Ford	UPS
R-CNN	87.16	49.59	**98.33**	86.76	68.55	80.38	70.11	76.59	88.47	84.49	**88.11**
OUR	86.11	50.11	94.82	**90.44**	64.45	**84.60**	71.25	76.55	89.56	85.09	85.98
R-CNN-BB	**88.52**	52.61	**98.33**	90.33	**71.16**	80.61	71.17	**76.72**	89.78	85.27	87.99
OUR-BB	88.03	**59.46**	95.19	90.19	67.52	81.83	**76.12**	72.73	**90.17**	**85.58**	86.93
Class	Stellaartois	Erdinger	Cocacola	Ferrari	Chimay	Texaco	Milka	Shell	Esso	bmw	mAP
R-CNN	**81.50**	52.92	67.02	88.94	64.81	**81.82**	58.66	72.73	89.92	82.07	74.07
OUR	80.29	50.54	66.73	89.74	66.41	80.68	54.08	72.49	90.15	79.60	73.96
R-CNN-BB	80.18	**70.65**	**72.22**	90.32	64.93	**81.82**	**62.21**	72.73	**97.93**	**82.46**	76.49
OUR-BB	79.72	61.24	67.43	**90.91**	**68.04**	79.80	61.58	**79.72**	89.66	78.05	**76.63**

Third, 33 softmax-layers were used in RCNN while ours only use 32 channels in softmax output. We focus on classifying different classes.

3 Sensitive Audio Information Detection on the Internet

Audio data is also a kind of inter-media for illegal information, through which, lawbreakers spread violence and religious extremism, like religious music, oath slogan and so on. Even identical audio context can have disparate voice properties for different individuals in various scenes. However, the melody information that music has, is identical even though individuals have unlike voice properties.

3.1 Humming Based Sensitive Audio Information Detection

The essence of Query by humming [10,11,13,16,17,22–24] is to detect a specific context of voice by utilizing these unchangeable melody information. In this paper, we put forward a new audio detection method which is based on melody feature. In this proposed method, Empirical Mode Decomposition(EMD) is introduced, with Dynamic Time Warping(DTW) combined. The whole framework is shown as Fig. 3.

The whole system can be loosely translated into three parts. First part focuses on dataset construction, in which various sensitive audio information is collected. And then, note feature and pitch feature are collected for each audio. Second part conducts pitch feature extraction of query data, after which feature transformation is applied to extract note feature. Third part is matching stage. Top N nearest neighborhoods with minimum EMD distance of note feature, are selected as candidates. Then, DTW is applied to these candidates to match distance of pitch feature. We re-rank candidates by linear weighting.

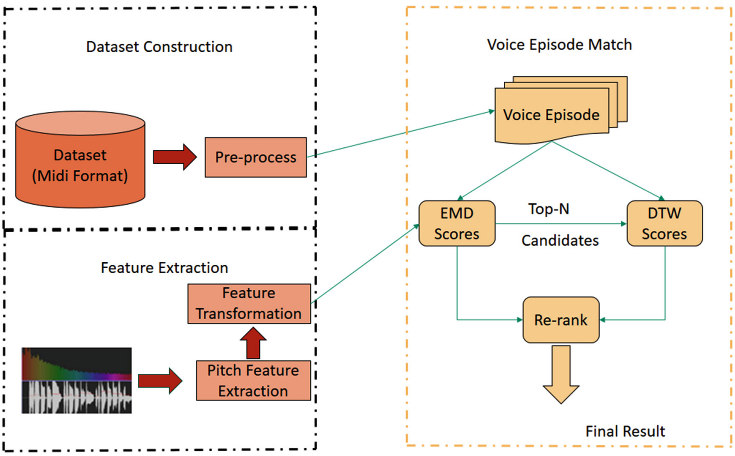

Fig. 3. A whole framework of sensitive audio detection.

3.2 Experiment and Analysis

To verify feasibility of our framework, we conduct simulation experiment on MIREX competition dataset, where a total of 2,048 songs exist, including 48 target humming songs and others belonging to noise data. Also, 4,431 humming songs are used as queries. Partial searching results are shown in Fig. 4. As we can see, the vast majority of humming query can find its corresponding source songs with a 93% retrieval rate.

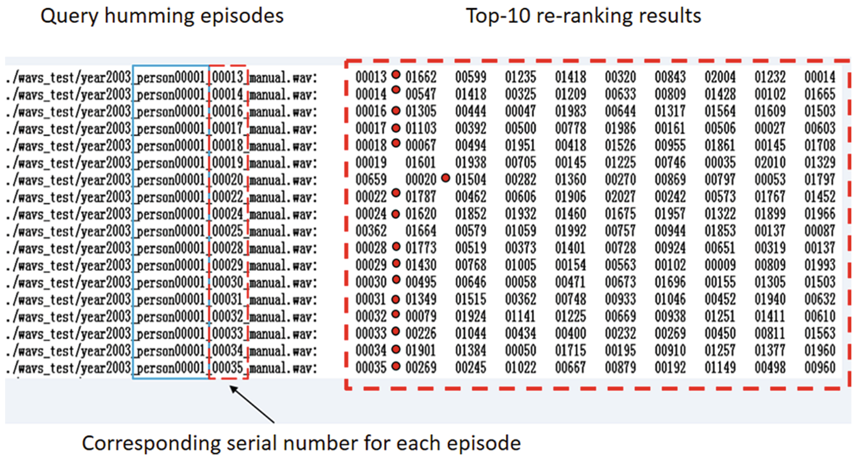

Fig. 4. A whole framework of sensitive audio detection.

4 Conclusion

Abnormal sensitive information on the Internet lies in various multimedia, like text, video or audio. As far as text type, existing algorithms can figure it out with efficient results and instantaneity. For video or audio, though enough works are insisting on them, they are still un-solved, which remains an open problem. In this paper, we propose two algorithms, i.e., iterative based semi-supervised deep learning model and Humming melody based search model, to detect abnormal visual and audio objects respectively. And experiments show the feasibility of our proposed methods.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
2. Cao, L., Luo, F., Chen, L., Sheng, Y., Wang, H., Wang, C., Ji, R.: Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. Pattern Recogn. **64**, 417–424 (2017)
3. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 109–122. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_8
4. Chen, X., Xiang, S., Liu, C.-L., Pan, C.-H.: Vehicle detection in satellite images by parallel deep convolutional neural networks. In: 2013 2nd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 181–185. IEEE (2013)
5. Cheng, X.Q., Jin, X., Wang, Y., Guo, J., Zhang, T., Li, G.: Survey on big data system and analytic technology. J. Softw. **25**(9), 1889–1908 (2014)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
7. Foote, J.T.: Content-based retrieval of music and audio. In: Voice, Video, and Data Communications, pp. 138–147. International Society for Optics and Photonics (1997)
8. Geronimo, D., Lopez, A., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1239–1258 (2010)
9. Ghias, A., Logan, J., Chamberlin, D., Smith, B.C.: Query by humming: musical information retrieval in an audio database. In: Proceedings of the Third ACM International Conference on Multimedia, pp. 231–236. ACM (1995)
10. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: ISMIR, vol. 2002, pp. 107–115 (2002)
11. Roger Jang, J.-S., Gao, M.-Y.: A query-by-singing system based on dynamic programming. In: Proceedings of International Workshop on Intelligent System Resolutions (8th Bellman Continuum), Hsinchu, pp. 85–89. Citeseer (2000)
12. Ji, R., Liu, W., Xie, X., Chen, Y., Luo, J.: Mobile social multimedia analytics in the big data era: An introduction to the special issue. ACM Trans. Intell. Syst. Technol. (TIST) **8**(3), 34 (2017)

13. Jiang, H., Xu, B., et al.: Query by humming via multiscale transportation distance in random query occurrence context. In: 2008 IEEE International Conference on Multimedia and Expo (2008)
14. Jiang, X., Ping, Y.: Research and implementation of lucene-based retrieval system of audio and video resources. Jisuanji Yingyong yu Ruanjian **28**(11), 245–248 (2011)
15. Liu, C.-C., Chang, P.-F.: An efficient audio fingerprint design for MP3 music. In: Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia, pp. 190–193. ACM (2011)
16. Liu, H., Ji, R., Wu, Y., Liu, W.: Towards optimal binary code learning via ordinal embedding. In: AAAI, pp. 1258–1265 (2016)
17. McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L., Cunningham, S.J.: Towards the digital music library: Tune retrieval from acoustic input. In: Proceedings of the first ACM International Conference on Digital Libraries, pp. 11–18. ACM (1996)
18. Su, S.-Z., Li, S.-Z., Chen, S.-Y., Cai, G.-R., Wu, Y.-D.: A survey on pedestrian detection. Dianzi Xuebao (Acta Electronica Sinica) **40**(4), 814–820 (2012)
19. Sun, Z., Bebis, G., Miller, R.: On-road vehicle detection: A review. IEEE Trans. Pattern Anal. Mach. Intell. **28**(5), 694–711 (2006)
20. Tian, Y., Luo, P., Wang, X., Tang, X.: Pedestrian detection aided by deep learning semantic tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5079–5087 (2015)
21. Uijlings, J.R.R., Van De Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vision **104**(2), 154–171 (2013)
22. Wang, L., Huang, S., Hu, S., Liang, J., Xu, B.: An effective and efficient method for query by humming system based on multi-similarity measurement fusion. In: International Conference on Audio, Language and Image Processing, ICALIP 2008, pp. 471–475. IEEE (2008)
23. Wang, Q., Guo, Z., Li, B., Liu, G., Guo, J.: Tempo variation based multilayer filters for query by humming. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 3034–3037. IEEE (2012)
24. Wu, X., Li, M., Liu, J., Yang, J., Yan, Y.: A top-down approach to melody match in pitch contour for query by humming. In: Proceedings of 5th International Symposium on Chinese Spoken Language Processing, Singapore (2006)
25. Yang, S., Luo, P., Loy, C.-C., Tang, X.: From facial parts responses to face detection: a deep learning approach. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3676–3684 (2015)
26. Zhong, B., Yuan, X., Ji, R., Yan, Y., Cui, Z., Hong, X., Chen, Y., Wang, T., Chen, D., Jiaxin, Y.: Structured partial least squares for simultaneous object tracking and segmentation. Neurocomputing **133**, 317–327 (2014)