COMPACT BILINEAR DEEP FEATURES FOR ENVIRONMENTAL SOUND RECOGNITION

Fatih Demir[1], *Abdulkadir Sengur*[1], *Hao Lu*[2], *Shahin Amiriparian*[3,4], *Nicholas Cummins*[3], *Björn Schuller*[3,5]

[1]Firat University, Technology Faculty, Electrical and Electronics Engineering Dept., Elazig, Turkey
[2]National Key Laboratory of Science and Technology on Multi-Spectral Information Processing,
School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[3]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[4] Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[5]GLAM – Group on Language, Audio, and Music, Imperial College London, UK
ksengur@firat.edu.tr

ABSTRACT

Environmental sound recognition (ESR) has extensive various civilian and military applications. Existing ESR methods generally tackle this problem by employing various signal processing and machine learning methods. Herein, an ESR paradigm based on feature extraction from pre-trained deep convolutional neural networks (CNN), the derivation of higher-order statistics by compact bilinear pooling and normalisation. In particular, we consider two deep ImageNet architectures for deep feature extraction, and the Random Maclaurin (RM) to produce the compact bilinear features. A support vector machine (SVM) with homogeneous mapping is used in the classification stage. Two publicly available environmental sound datasets are used to verify the efficacy of the approach namely, ESC-50 and ESC-10. We compare the proposed method with various previous state-of-the-art methods. Presented results indicate the suitability of the higher-order statistics of DEEP SPECTRUM representations for ESR classification tasks.

Index Terms**—**Environmental sound classification, deep spectrum features, convolutional neural networks, compact bilinear pooling

1. INTRODUCTION

Environmental sound recognition(ESR) is an important computer audition topic, which helps achieve content-based sound retrieval [1], smart home monitoring for elderly people [2], improving autonomous navigation [3], surveillance [4], and sound-based animal and bird species determination [5], to name but a few tasks. As with many other areas of computer audition, deep neural networks (DNN), in particular convolutional neural networks (CNN), have become the predominant approach for ESC. This is due in part to the ability of CNNs to learn robust, task specific feature representations.

CNNs generally consist of a combination of convolutional, pooling, and fully connected layers. Feature representation are learnt through thefiltering action of convolutional layers. Work of presented by Piczak [6], demonstrated that a CNN, comprising of two convolutional layers with max pooling and two fully connected layers, was easily able to outperfrom a Mel-spectrum baseline system for a range of ESC tasks. A range of neural network based approaches was compared in [7]. The presented results indicate the suitability of using CNNs to learn features directly from the mel-spectrum for ESC tasks. Recently, Aytar et al. [8] proposed the Soundnet system for ESC. This approach which transfers discriminative visual knowledge into the audio modality, can be considered state-of-the-art for ESC.

Herein, we propose and develop an ESR approached based on compact bilinear deep spectrum based CNN features. Bilinear CNN features have been shown to be effective infine-grained image classification tasks [9], however, they produce high dimensional feature vectors [10]. In this regard, the*compact bilinear model*has been proposed to alleviate this drawback [10]. The proposed method uses pre-trained CNN models, namely VGG-M and VGG-D for DEEP SPECTRUM feature extraction [11]. DEEP SPECTRUM features,first introduced by Amiriparian et al. [12], have shown promise in a range of computer audition tasks such as acoustic surveillance [13] and speech-based emotion recognition [14]. The *Random Maclaurin*(RM) approximation is then utilised to obtain the compact bilinear features [10, 15]; RM is considered due to its computational efficiency [10]. The approach is tested on two publicly available environmental sound datasets namely, ESC10 and ESC50 [16], with the proposed method achieving state-of-the-art results on both datasets.

The rest of the paper is organized as follows. Related works are discussed in Section 2.The proposed methodology is then introduced in Section 3. The experimental details and results are presented in Section 4. We conclude the paper and offer future work directions in Section 5.

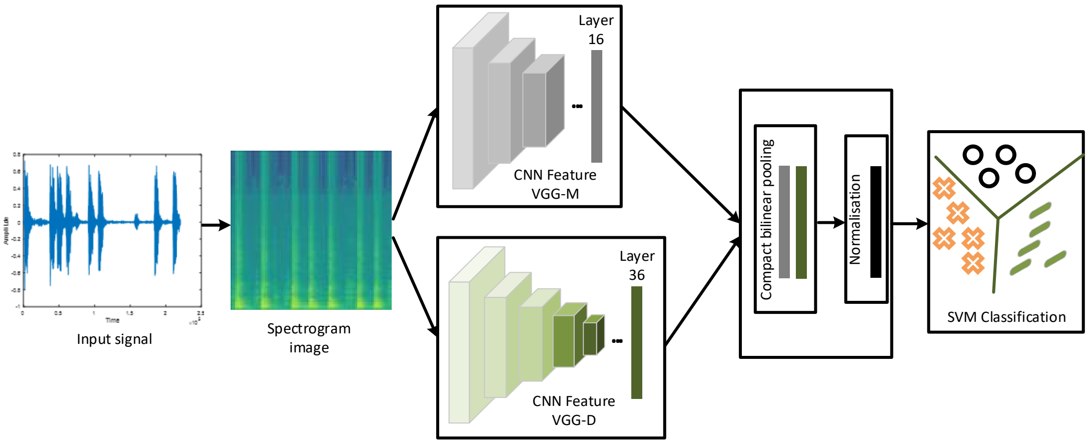

Fig. 1. An illustrative overview of the proposed methodology. The input audio signal () is converted into a spectrogram from which deep features are extracted using both the VGG-M and VGG-D deep CNN imageNets. Compact bilinear pooling and normalisation is then used to form the final feature representation for classification with a support vector machine. The displayed audiofile is dog bark File ID: 1-59513-A

2. RELATED WORKS

Neural Networks, CNNs in particular, are now the predominant approach for ESR [17]. A limiting factor when using CNNs for ESC is the amount of available training data. In this regard, Salamon et al. [18] explored data augmentation approaches for CNN-based ESC. Results presented by the authors indicate that data augmentation enabled the training of a deeper CNN model capable of state-of-the-art performances on the UrbanSound8K dataset [19]. Similarly, to locate salient sound events in the UrbanSound8K dataset, Su et al. [20] proposed a weakly supervised learning approach based on CNN. The proposed approach was able to perform accurate sound event localisation, without specific training using temporal annotations.

In image processing it is now becoming standard practice to use deep pre-trained CNNs, such as AlexNet [21], VGG16 and textscVGG19 [11], for feature extraction [22]. The use of pre-trained image CNNs for feature extracted has also transitioned into the audio domain; as previously mentioned, DEEP SPECTRUM features have been used for a wide range of audio based detection tasks [12, 13, 14]. Recently, Ren et al. [23] explored VGG16 based DEEP SPECTRUM derived from either spectrograms and scalograms for the task of acoustic scene classification. highlighting the promise of this approach, a fusion of three different DEEP SPECTRUM approaches was able to outperform the Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 challenge baseline system.

The DEEP SPECTRUM feature extraction methodology produces a high-dimensional and sparse feature representation. Compact Bilinear Pooling [10] represents a promising technique to help alleviate this issue. Compact Bilinear Pooling has been shown to produce state-of-the-art results in image classification tasks such as fine-grained visual recognition [9, 10], texture classification [10] and indoor scene recognition [10]. To the best of the authors knowledge, this is the first time they have been explored for ESC.

3. PROPOSED METHOD

The proposed approach is based on the processing of DEEP SPECTRUM features with compact bilinear pooling and normalisation (cf. Figure 1). The following subsections focus on the bilinear pooling (cf. Section 3.1) and compact bilinear pooling (cf. Section 3.2) operations which are the novel extensions to the DEEP SPECTRUM paradigm. For further information on DEEP SPECTRUM feature extraction the interested reader is referred to both [12, 14].

3.1. Bilinear Pooling

Bilinear pooling or second-order pooling was first introduced to the computer vision community in [24]. It constructs a global feature vector for a given input image using the operation:

$$B(X) = \sum_{s\in S} x_s x_s^T,(1)$$

in which$B(X)$is a$c\times c$matrix often denoted asc^2 vector,X= $(x_1, ..., x_{|S|}, x_s \in \mathbb{R}^c)$is a set of local features in our case Deep Spectrum Features, and$S$a set of row and column spatial locations. If the local features are ex-

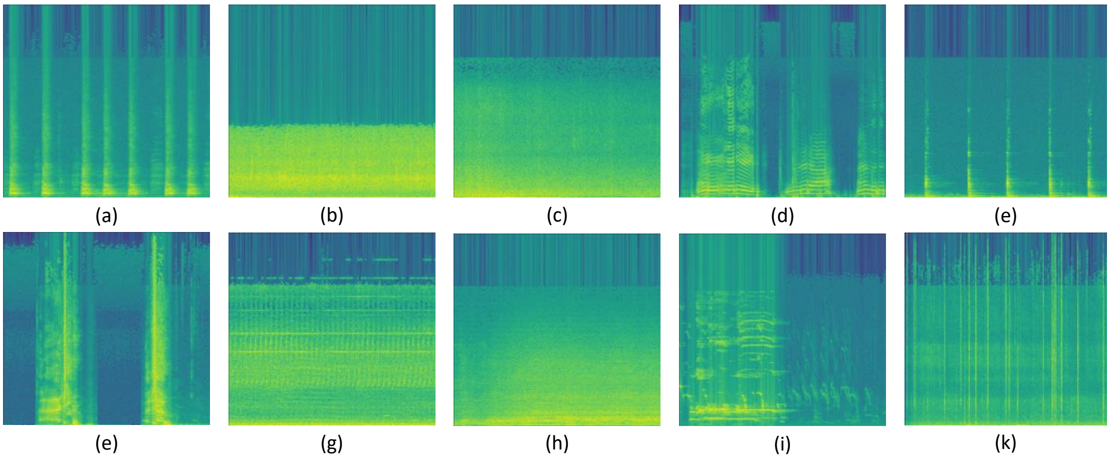

Fig. 2. Exemplar spectrogram images taken from each of 10 classes in the ESC-10 dataset: a) dog bark (File ID: 1-59513-A), b) rain (File ID: 1-17367-A), c) sea waves (File ID: 1-28135-A), d) a baby cry (File ID: 1-22694-A), e) a clock ticking (File ID: 1-21934-A), f) a person sneezing (File ID: 3-144692-A), g) a helicopter (File ID: 1-172649-A), h) a chainsaw (File ID:1-19898-A), i) a rooster (File ID: 3-116135-A), and j) afire crackling (File ID: 1-4211-A).

tracted from a CNN model, the obtained $B(X)$ feature matrix is known as bilinear CNN features.

3.2. Compact Bilinear Pooling

As bilinear pooling constructs a high dimensional feature space, a low dimensional projection function, $\phi(x) \in \mathbb{R}^d$ where $d \ll c$, is needed to create a more compact feature representation [10]. Given two sets of local features X, Y and a comparison kernel $k(x, y)$ which satisfies $\langle\phi(x), \phi(y)\rangle \approx k(x, y)$ the compact bilinear pooling operation is given by:

$$\begin{aligned} B(X) &= \sum_{s\in S}\sum_{u\in U}\langle\phi(x_s),\phi(y_u)\rangle^2 \\ &\approx \sum_{s\in S}\sum_{u\in U}\langle\phi(x),\phi(y)\rangle \\ &\equiv \langle C(X), C(Y)\rangle, \end{aligned} \tag{2}$$

where $\langle\cdot,\cdot\rangle^2$ denotes a second order polynomial kernel and

$$C(X) = \sum_{s\in S}\phi(x_s) \tag{3}$$

the compact bilinear feature vector. In this work, the Random Maclaurin (RM) approximation is used as the low dimensional projection function. In the RM procedure $\omega_1, \omega_2 \in \mathbb{R}^c$ are two random $-1, +1$ vectors and $\phi(x) = \langle\omega_1, x\rangle\,\langle\omega_2, x\rangle$, then for non-random vectors $x, y \in \mathbb{R}^c$:

$$\begin{aligned} E[\phi(x),\phi(y)] &= E[\langle\omega_1, x\rangle\,\langle\omega_2, y\rangle]^2 \\ &= \langle x, y\rangle^2. \end{aligned} \tag{4}$$

Therefore, when using RM, each projected entry requires the expectation of the quantity to be approximated. For full details on this procedure the reader is referred to [10, 15]. After parametrisation of the compact features, we apply a signed squared root operation and instance-wise L_2 normalisation before classification.

4. DATASETS AND EXPERIMENTAL WORKS

4.1. Datasets

The ESC-10 dataset contains 400 audiofiles in 10 sound event categories: dog bark, rain, sea waves, baby cry, clock tick, person sneeze, helicopter, chainsaw, rooster, andfire crackling [16]. The ESC-50 dataset is an expanded corpora containing 2 000 short audiofiles in 50 classes[16]. Each ESC-50 class contains 40 audiofiles in 5 main categories. These categories include animals, natural soundscapes and water sounds, human (non-speech) sounds, interior-domestic sounds, and exterior-urban noises, respectively. All audio files in the ESC-10 and ESC-50 datasets are 5 seconds in duration and have a sampling frequency of 44.1 kHz.

4.2. Experimental Settings

All experimental work is performed using the Matlab software package 2017b on a computer having an Intel Core i7-4810 CPU and 32 GB memory. To form the spectrogram images a Hamming window of width 1 024 ms and overlap 256 ms is used with the number of Fast Fourier Transform (FFT) points set to 1 024; these values were set empirically during initial experimentation. The power spectral density of the spectrogram images are then computed on the dB power scale and are saved with the *viridis* colour map noting that this colour map has previously shown to be highly suitable

Table 1. Obtained accuracy values for each compact bilinear model.

	Accuracy (%)	
Models	ESC-10	ESC-50
VGG-M & VGG-M	89.0	70.4
VGG-D & VGG-D	86.0	65.8
VGG-M & VGG-D	**92.5**	**74.6**

Table 2. Performance comparison of the proposed method with state-of-the-art methods.

	Accuracy (%)	
Models	ESC-10	ESC-50
CNN (Piczak) [6]	81.1	64.5
SoundNet [8]	92.2	74.2
Proposed Method	**92.5**	**74.6**

when extracting DEEP SPECTRUM images [12]. The initial spectrogram images have a size of 875×656 and are resized to 224×224 to be compatible with the VGG nets. Exemplar spectrograms from each of the 10 classes in ESC-10 are given in Figure 2.

We extracted 4 096-dimensional DEEP SPECTRUM feature vectors from the spectrogram images by using the activations of layer 16 and layer 36 of VGG-M and VGG-D models respectively [11]. It is worth noting that initial experiments were also conducted using other layers, namely layer 18 of VGG-M and layer 34 of VGG-D, however using the activations from these layers yielded weaker performances. For the compact bilinear pooling of the extracted feature vectors, three possible combinations of the extracted feature sets are considered during the experimental works and their achievements are given in Table 1 for both datasets. In thefirst model, we use the VGG-M architecture for both feature vectors. Similarly for the second model, we use the VGG-D for both feature vectors. Finally, in the third model, both VGG-M and VGG-D architectures are used to obtain the compact bilinear feature vectors.

A support vector machine (SVM) classifier is used to determine the class label of the input signal. A SVM was considered due to its well established ability to handle sparse data representations and its robustness to smaller amounts of training data. Specifically, we implemented a SVM classifier with homogenous mapping from the LIBLINEAR library with the L2-regularised L2-loss dual solver [25]. The SVM cost function parameter$C$was searched in the range of$[10^{-4}, 10^{-3}, \cdots, 10^{3}]$.

4.3. Results

When comparing between our three different approaches, the highest accuracy is obtained with the third compact bilinear model (VGG-M & VGG-D) where 92.50 % (ESC-10) and 76.1 % (ESC-50) accuracies were obtained (cf. Table 1).The second best accuracy values are obtained for the second compact bilinear model (VGG-D & VGG-D), achieving accuracies of 90.5 % and 73.0 % for the ESC-10 and ESC-50 datasets respectively. The weakest classification results were produced by thefirst model (VGG-M & VGG-M) which achieved ESC-10 and ESC-50 accuracies of 88.5 % and 72.2 %, respectively.

It is possible that the second models accuracy is higher than thefirst one's due to VGG-D having a deeper architecture than VGG-M [11]. In addition, the third models strong result is reasonable as it combines DEEP SPECTRUM from both VGG models to produce thefinal compact bilinear feature vector. These results indicate that the VGG-M and VGG-D captures different characteristics of input signals. By synthesizing both information via bilinear pooling, we can better make use of the complementarity nature of both feature representations.

Further, we compare the obtained accuracies with previous aforementioned state-of-the-art method (cf. Table 2); namely the CNN system proposed by Piczak [6] and the SoundNet system [8]. Highlighting the advantages of the proposed method, our third set-up (VGG-M & VGG-D) outperformed Piczak's CNN system and matched performance with SoundNet, on both datasets. This result highlights that the pooled higher-order statistics of DEEP SPECTRUM features capture important ESR information.

5. CONCLUSION

In this paper, we proposed a compact bilinear pooling method for environmental sound recognition (ESR). The proposed method utilises compact bilinear pooling to effectively exploit higher-order statistics of CNN-based DEEP SPECTRUM features. Using two publicly available datasets are used in experiments, and various pooling models, based on the pretrain VGG neural nets, were compared to assess the effectiveness of the proposed approach. In addition, our approach was compared against various state-of-the-methods. The obtained results indicate the effectiveness of compact bilinear pooling for ESR tasks. In the future work, we aim to investigate multimodal bilinear pooling schemes for ESR.

6. ACKNOWLEDGEMENTS

This work was supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreement No. 338164 (ERC StG iHEARu).

7. REFERENCES

[1] S. Duan, J. Zhang, and M. Roe, P.and Towsey, "A survey of tagging techniques for music, speech and environmental sound,"*Artificial Intelligence Review*, vol. 42, no. 4, pp. 637–661, Dec. 2014.

[2] J. C. Wang, H. P. Lee, J. F. Wang, and C. B. Lin, "Robust Environmental Sound Recognition for Home Automation,"*IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 1, pp. 25–31, Jan. 2008.

[3] S. Chu, S. Narayanan, C. c. J. Kuo, and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," in*2006 IEEE International Conference on Multimedia and Expo (ICME)*, Toronto, Canada, July 2006, pp. 885–888, IEEE.

[4] M. Cristani, M. Bicego, and V. Murino, "Audio-Visual Event Recognition in Surveillance Video Sequences,"*IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, Feb. 2007.

[5] F. Weninger and B. Schuller, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in*2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 337–340, IEEE.

[6] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in*2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, Sep. 2015, pp. 1–6, IEEE.

[7] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of Deep Learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 126–130, IEEE.

[8] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning Sound Representations from Unlabeled Video," in*Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 892–900. Curran Associates, Inc., 2016.

[9] T-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN Models for Fine-Grained Visual Recognition," in*The IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1449–1457, IEEE.

[10] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact Bilinear Pooling," in*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, Jun. 2016, pp. 317–326, IEEE.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition,"*CoRR*, vol. abs/1409.1556, 2014.

[12] S. Amiriparian, M. Gerczuk, S. Ottl, N.s Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in*Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, Aug. 2017, pp. 3512–3516, ISCA.

[13] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis," in *Proceedings 31st International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, July 2018, IEEE, IEEE, 8 pages, to appear.

[14] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in*Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, CA, Oct. 2017, MM '17, pp. 478–484, ACM.

[15] P. Kar and H. Karnick, "Random Feature Maps for Dot Product Kernels," in*Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Neil D. Lawrence and Mark Girolami, Eds., La Palma, Canary Islands, Apr. 2012, vol. 22 of*Proceedings of Machine Learning Research*, pp. 583–591, PMLR.

[16] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in*Proceedings of the 23rd ACM International Conference on Multimedia*. Oct. 2015, MM '15, pp. 1015–1018, ACM.

[17] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge,"*IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2018.

[18] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification,"*IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[19] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in*Proceedings of the 22Nd ACM International Conference on Multimedia*. Oct. 2014, MM '14, pp. 1041–1044, ACM.

[20] T. W. Su, J. Y. Liu, and Y. H. Yang, "Weakly-supervised audio event detection using event-specific Gaussianfilters and fully convolutional networks," in*2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, Mar. 2017, pp. 791–795.

[21] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in*Proceedings of the 26th Annual Conference on Advances in Neural Information Processing Systems*, Lake Tahoe, NV, 2012, pp. 1097–1105, NIPS.

[22] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in*2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, Jun. 2014, pp. 512–519, IEEE.

[23] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in*Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE '17)*, Munich, Germany, Dec. 2017, pp. 113–117, IEEE.

[24] J. B. Tenenbaum and W. T. Freeman, "Separating Style and Content with Bilinear Models,"*Neural Computation*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000.

[25] R-E. Fan, K-W. Chang, C-J. Hsieh, X-R. Wang, and C-J. Lin, "LIBLINEAR: A library for large linear classification,"*Journal of Machine Learning Research*, vol. 9, no. Aug, pp. 1871–1874, 2008.