IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

On the Discriminative Power of Learned vs. Hand-Crafted Features for Crowd Density Analysis

Mohamed Amine Marnissi
Université de Sousse,
Ecole Nationale d'Ingénieurs de Sousse,
LATIS- Laboratory of Advanced Technology
and Intelligent Systems, 4023, Sousse, Tunisie
medamine.marnissi@eniso.u-sousse.tn

Hajer Fradi
Université de Sousse,
Institut Supérieur des Sciences Appliquées et de Technologie de Sousse,
LATIS- Laboratory of Advanced Technology
and Intelligent Systems, 4023, Sousse, Tunisie
hajer.fradi@issatso.rnu.tn

Jean-Luc Dugelay
Digital Security Department
EURECOM
Sophia Antipolis, France
jean-luc.dugelay@eurecom.fr

***Abstract*—Crowd density analysis is a crucial component in video surveillance mainly for security monitoring. This paper proposes a novel approach for crowd density classification, in which learned features substitute the commonly used hand-crafted features. In particular, the approach consists of employing deep networks to extract useful crowd features that can further be manageable by a classifier. This process is favorable for crowd features extraction due to the large learning capability of deep networks compared to traditional methods based on hand-crafted features. The proposed approach is evaluated on three challenging datasets, and the results demonstrate the effectiveness of learned features for crowd density classification. Furthermore, we include an extensive comparative study between different learned/hand-crafted features in order to investigate their discriminative power to handle such problems. Their performance is evaluated using different classifiers and strategies as well.**

I. INTRODUCTION

Studying crowd phenomenon is becoming of significant interest with the increasing number of popular events that gather many persons such as in subways, religious events, public demonstrations, sport events and car traffic. In this context, crowd analysis has emerged as a major topic for crowd monitoring and management in visual surveillance community [1]–[3]. In particular, the estimation of crowd density is receiving much attention for security reasons. It could be used for developing crowd management strategies by measuring the comfort level in public spaces. Also, its automatic monitoring is important to prevent overcrowd which can potentially lead to disastrous and fatal accidents. For these reasons, early detection of unusual situations in large scale crowd is required and appropriate decisions for safety control have to be taken to insure assistance and emergency contingency plan.

In this context, many recent works in the field of automatic video surveillance have been proposed to address the problem of crowd density analysis [4]–[6]. Precisely, significant progress has been made in this field over the last decade using low level and holistic features. This paradigm is proposed as an alternative solution to pedestrian detection based methods (such as [7]) because of the partial occlusions that often occur in the crowd, and that make delineating people a challenging task. Thus, recent works mostly bypass the task of detecting people and instead focus on learning a mapping between a set of low level features and the crowd density.

The taxonomy of crowd density analysis methods can be categorized into two groups: crowd counting and crowd density classification for which the goal is to estimate the number of people, or to alternatively estimate the crowd level [5], [6]. In this paper, we particularly emphasize the need for estimating crowd level mainly to prevent overcrowd when the number of persons flooding some areas exceeds a certain level (e.g. in some religious or sport events). According to the classification introduced in [8], the crowd density can be categorized into 5 levels: free, restricted, dense, very dense, and jammed flow.

One of the key aspects of crowd density analysis is the features extraction step. Early attempts to handle this problem generally made use of texture features which are more frequently used than statistical features (usually employed for people counting problem). Based on the observation that regions of low density crowd tend to present less dense features compared to a high-density crowd, many texture features such as: Gray Level Co-occurrence Matrix (GLCM) [9], Gradient Orientation Co-occurrence Matrix (GOCM) [10], Gray Level Dependence Matrix (GLDM) [11], Gabor filter [12], and dynamic texture features [13] have been proposed so far to handle the problem of crowd level classification.

The use of Local Binary Pattern (LBP) [14] as local texture features has been an active topic in this field. In [15], LBP is used in blocks, from which Dual-Histogram LBP (DH-LBP) is computed and K-means clustering is used for crowd density classification. In [16], the dynamic texture of the walking crowd is extracted using a sparse spatio-temporal local binary pattern (SST-LBP) features. Afterwards, a statistical

978-1-7281-2009-6/$31.00 paper N-20479.pdf

property of SST-LBP is employed as crowd features and then classified into a range of density levels by adopting Support Vector Machine (SVM). In [17], an histogram model based on improved uniform LBP that has the advantages of being invariant to intensity and rotation is proposed. In the same context, a combination of principal component analysis (PCA) and linear discriminant analysis (LDA) to project high-dimensional LBP to low-dimensional discriminative feature space has been proposed in [18].

Some related works to crowd density estimation using LBP descriptor include other texture features to encode image characteristics. For instance, in [19], GLCM is computed on LBP image instead of the original gray image. The resulting texture descriptor called LBP Co-occurrence Matrix (LBPCM) is constructed from several overlapping cells in an image block, and then classified into different crowd density levels. Another fusion approach has been proposed in [12], where Uniform LBP features that reduce the dimension of the conventional LBP and Gabor features that are extracted after convolving the original image with a bank of Log-Gabor filters at different scales and orientations are combined in a single feature vector to train a multi-class SVM classifier.

In this paper, we particularly intend to investigate the discriminative power of such hand-crafted features, essentially based on LBP descriptor since it is the most commonly used in the field of crowd density analysis compared to learned features. Indeed, deep learning models [20] have recently shown good performance in different applications, essentially for image classification and recognition by means of Convolutional Neural Networks. Deep feature representation in these networks can act as a set of feature extractors which have the potential of being representative and generic enough with the increasing depth of layers [21]. Similar attempts have been investigated in other applications such as gender recognition [22], pedesterian and face dectecion in [23].

The increasing interest in deep learning techniques has been expanded to crowd people counting problem [24]–[27]. For instance, a Multi-column Convolutional Neural Network (MCNN) architecture is proposed in [24] to estimate the crowd number in a single image from almost any perspective. Also, a modified three-tier MCNN architecture is employed in [25]. In [26], the authors propose to consider the temporal information in video sequences by using a variant of convolutional LSTM (ConvLSTM) for crowd counting. Even though many research works have been conducted using deep networks for crowd counting, fewer studies for crowd density classification problem have been explored in this context, except [28], where a cascade optimized convolutional neural network based on the multi-stage ConvNet for crowd density estimation is adopted. This problem will be exceedingly studied in the current paper, where we are essentially interested in investigating the discriminative power of learned features vs. hand-crafted features for exhibiting relevant crowd features. To achieve this goal, a comparative study between different learned feature representations namely, by pre-trained models and a proposed Convolutional Neural Network (CNN) is presented. Furthermore, the performance of learned features is compared to hand-crafted features. The results using different classifiers are compared as well.

The contribution of this paper is three-fold: First, we substitute the commonly used hand-crafted features for crowd level classification with learned features by adopting various deep networks, such as pre-trained models and a proposed CNN architecture that improves the overall classification rate using three challenging datasets. Second, we conduct extensive experiments using various classifiers to evaluate both of learned and hand-crafted features in order to prove the effectiveness of our proposal. Third, additional tests are provided and analyzed to demonstrate the generalization ability and the representative capacity of learned features regarding hand-crafted features.

The remainder of the paper is organized as follows: In Section II, different techniques of feature extraction (both of hand-crafted and learned features) for crowd density classification are presented. The different presented feature extractors are evaluated using three datasets and the experimental results are analyzed in Section III. Finally, we conclude and present some potential future works in Section IV.

II. Feature extraction for crowd density classification

To handle the problem of crowd level classification, the feature extraction is a key step as in any classification problem. In this section, different proposed techniques for feature extractors are presented. Generally, existing feature representations can be divided into two categories: the hand-crafted features and the learned features. The hand-crafted features are those extracted from separate images according to some predefined algorithms based on the expert knowledge. The learned features are contrary derived from a dataset by training [21]. In this paper, we consider 2 types of learned features: those extracted from pre-trained models and by training a CNN architecture. For hand crafted features, LBP descriptor is considered since it is the most commonly used in the field of crowd density analysis.

A. Hand-Crafted Features

Based on the observation that high density crowd has fine patterns of texture, and images of low density have coarse patterns of texture, texture features can be employed for crowd level classification. Major attempts to handle this problem made use of LBP [14] as hand-crafted features. This descriptor has aroused increasing interest in many applications of computer vision field, in particular, it has been extensively studied in face recognition. Likewise, significant progress has been made in the field of crowd density analysis using this descriptor. The advantage of using LBP as feature extractor is that it is a powerful descriptor that embeds the structure of the local image texture which is highly relevant to the crowd density.

LBP operator is based on labeling the pixels of an image by thresholding the 3 x 3 neighborhood of each pixel with the center value and considering the result as a binary digit. Then,

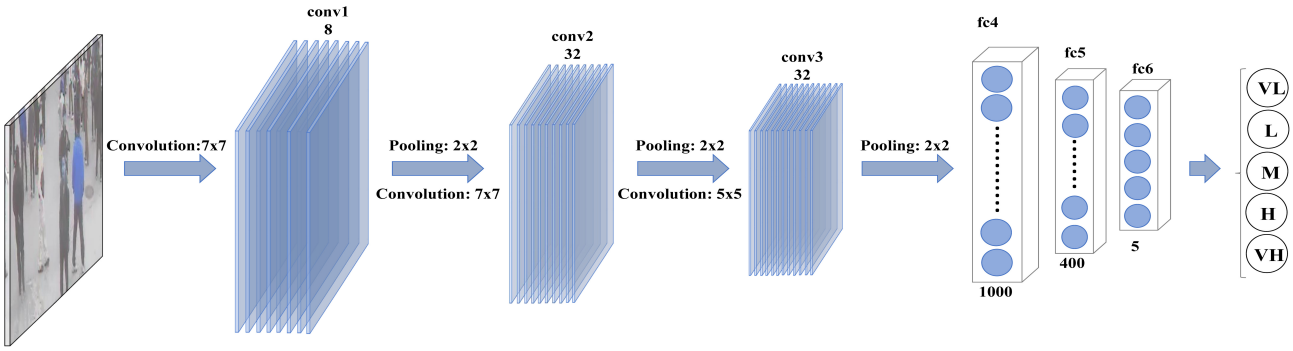

Fig. 1. The proposed crowd CNN architecture composed of three convolutional layers and three fully connected layers for crowd level classification: 5 classes of density: Very Low (VL), Low (L), Medium(M), High(H), and Very High(VH)

a binary number is obtained by concatenating all binary values in a clockwise direction, starting from the top left neighbor. Thus, for a given pixel (x_c, y_c), the LBP code in decimal form is defined as:

$$LBP(x_c, y_c) = \sum_{p=0}^{P-1} S(i_p - i_c)2^p$$

where i_c and i_p denote, respectively, the gray values of the center pixel and the P surrounding pixels. S refers to a thresholding function defined as: $S(x) = \begin{cases} 1 & \text{if } (x \geq 0) \\ 0 & \text{otherwise} \end{cases}$

In this paper, from each image under analysis, the histogram sequence is employed by computing the occurrence of LBP codes. To justify the choice of LBP as hand-crafted features, other texture features are considered for comparisons: GLCM, Gabor and a combination of Uniform LBP and Gabor filter as recently proposed in [12]. For GLCM features, 4 statistical properties are extracted: contrast, homogeneity, energy and entropy. For Gabor, two types of features are computed: local energy and mean amplitude using a bank of Log-Gabor filters at five different scales and six different orientations. These features are combined in [12] with Uniform LBP which is a reduced version of the conventional LBP of size 59 which results in a feature vector of size 119 in total.

B. Learned Features

Deep learning has reformed machine learning field and has consequently brought a revolution to the computer vision commmunity in recent years. Different algorithms have been proposed so far, marked by a cascade of many layers organized in a hierarchical way, where each layer adds certain abstraction to the overall feature representation [21]. The deep networks considered in this paper, are based on Convolutional Neural Networks (CNN). Since the final convolutional layers of a CNN architecture encode high level features, it is possible to consider these deep layers for features extraction. Precisely, in this paper, to extract learned features, we propose a CNN architecture. Also, the most commonly known pre-trained CNN models are evaluated.

1) Learned features by a proposed CNN architecture: Since CNN architectures embed different representations at different levels of abstraction, the deep layers can be employed as features extractor. Under this perspective, we propose a simple CNN architecture as shown in Fig. 1 which contains three convolutional and three fully connected layers. This architecture is referred as CrowdCNN in the rest of the paper. It is basically inspired from a CNN model in [29] which is initially proposed to handle the problems of people counting and density map estimation. We made some modifications to this architecture mainly in the last two layers to adapt it to the problem of crowd classification. Also, we changed the number of filters in each layer while keeping the same size of filters following the Bayesian optimization to find the optimal network parameters and the training options [30].

As depicted in Fig. 1, the first convolutional layer has 8 7x7x3 filters, the second convolutional layer has 32 7x7x32 filters and the last convolutional layer has 32 5x5x32 filters. Each convolutional layer is followed by a 2x2 max pooling layer with 2x2 stride. After each convolution layer, Rectified Linear Unit (ReLU) as activation function is applied. The proposed network is ended by three fully connected layers. The fc4 and fc5 have 1000 and 400 neurons, respectively. The last layer fc6 has 5 neurons which is the number of classes.

2) Learned features by pre-trained CNN models: The features extraction step can be performed using pre-trained models. Fine-tuned high-level CNN features have shown good performance for many applications [23], [31], [32]. Typically, a CNN model previously trained on a large dataset for a given classification task can be fine-tuned for another classification task. The high-level features of a CNN model can be retrieved from high layers and can be further fed into a classifier since higher layers are less dependent on the dataset compared to lower layers. By adopting this approach of training, we take advantage of the large-scale training data of CNN models. After fine-tuning a pre-trained model, the weights from one fully connected layer are used for classification.

In this paper, we compare different popular CNN pre-trained models for features extraction, namely, AlexNet [33], VGG-19 [34] and Inception-v3 [35]. For more details, AlexNet

was proposed for the first time in ImageNet ILSVRC-2012 competition and won the challenge with a significant margin compared to the second-best entry. The network consists of 11x11, 5x5, 3x3 convolutional layers, max pooling, and ReLU activations. It employs dropout to reduce overfitting in the fully-connected layers. More details about AlexNet architecture can be found in [33]. VGG-19 [34] is one of the best performing ConvNet models that achieved 7.3% error rate in the ILSVRC-2014 classification challenge. It consists of 19 layers with very small convolutional filters (of size 3x3) and is very appealing thanks to its uniform architecture. Inception-v3 [35] is the third pre-trained model that we employ for crowd features extraction. Its architecture incorporated all precedent upgrades of inception architectures. It utilized in addition RMSProp optimizer, factorized convolutions, additional regularization with batch-normalized auxiliary classifiers and label-smoothing to scale up the network. This architecture has achieved good results on ILSVRC-2012 classification benchmark with 3.5% as top-5 error rate and 17.3% as top-1 error rate on the validation set.

III. Experimental results

A. Datasets and Experiments

The proposed features are evaluated within different challenging crowded scenes from multiple datasets. In particular, we test hand-crafted and learned features on three challenging datasets: PETS 2009 [36], MALL [37] and HNUCROWD [38]. For the three datasets, different crowd levels are defined according to the range of people in the scene (as specified in [38]), see Table I.

Crowd Level	Label	Range of people
Very low	1	[0,10)
Low	2	[10,20)
Medium	3	[20,30)
High	4	[30,40)
Very high	5	≥ 40

TABLE I
Definition of different crowd levels according to the range of people.

PETS dataset [36] is a widely used dataset for different video surveillance applications, mainly our experiments are performed on section S_1, originally dedicated to assess person count and density estimation algorithms. Two other videos from sections S_2 and S_3 are employed to reach the fourth level (High) of the crowd. Therefore, using this dataset, only four levels of the crowd density are experimented. For each crowd level, 200 frames are selected. MALL [37], is a publically available dataset collected from an accessible surveillance camera in a shopping mall, extensively used for crowd counting. This dataset contains 2000 annotated frames of moving and stopping pedestrians with different lighting conditions. Using this dataset and according to the defined crowd levels in Table I, only frames for the three last crowd levels (corresponding to high density scenes) are available. HNUCROWD [38] is another dataset used in our experiments which enables us to experiment the five levels of crowd. This dataset was captured from the closed circuit television (CCTV) surveillance system of Hebei Normal University in China. It is a challenging dataset since some images contain disruptors such as moving cars, which are common scenarios in real scenes.

More details about the three datasets are given in Table II. Fig. 2 shows some sample images of different crowd levels from the three aforementioned datasets.

Crowd Level	Number of images		
	PETS 2009	MALL	HNUCROWD
Very low	200	-	300
Low	200	-	300
Medium	200	260	300
High	200	260	300
Very high	-	260	300

TABLE II
Details of number of frames used in the experiments from the three datasets: PETS 2009, MALL and HNUCROWD.

As described in Section II, both of hand-crafted and learned features are evaluated for crowd level classification. This multi-classification problem is 4-class, 3-class and 5-class for PETS, MALL and HNUCROWD datasets, respectively as depicted in the previous table and Fig. 2. Each time the overall available frames are randomly split into training and testing sets with 60% of the data as training set. For instance, this results in a 4-class training set of 120 frames and a testing set of 80 frames on PETS dataset. For tests, the feature vector is identified as one of the classes by the multi-class SVM classifier following one-vs-one strategy. The top-1 identification accuracy is reported for hand-crafted features and compared to learned features in order to demonstrate the discriminative power of each category of features. Furthermore, within each category, different methods are investigated, for instance for hand-crafted features, LBP performance is compared to 3 other texture features: GLCM, Gabor and Uniform LBP+Gabor [12]. Likewise, for learned features, extensive comparative study is given in order to highlight the effectiveness of the proposed architecture and the pre-trained models. Whatever the adopted feature extractor is, its performance is evaluated using SVM classifier (using both of linear and RBF kernels). Comparisons to other frequently used classifiers, namely decision tree [39], Bagging predictors [40], KNN [41], and subspace KNN [42] are also provided.

B. Results of hand-crafted features and analysis

We first report the classification accuracy obtained by applying SVM classifier using linear and RBF kernels on LBP

Fig. 2. Sample frames from the three experimented datasets: From top to bottom: HNUCROWD, MALL and PETS 2009 crowd datasets. From left to right: 5 levels of crowd density: free flow (very low), restricted flow (low), dense flow (medium), very dense flow (high) and jammed flow (very high)

Dataset	Classifier	LBP features
HNUCROWD	Linear SVM	88.33
	RBF SVM	**92.5**
MALL	Linear SVM	77.56
	RBF SVM	**85.58**
PETS	Linear SVM	78.44
	RBF SVM	**91.88**

TABLE III
THE CLASSIFICATION ACCURACY USING LBP FEATURES BY EMPLOYING LINEAR AND RBF KERNELS OF SVM ON THE THREE EXPERIMENTED DATASETS.

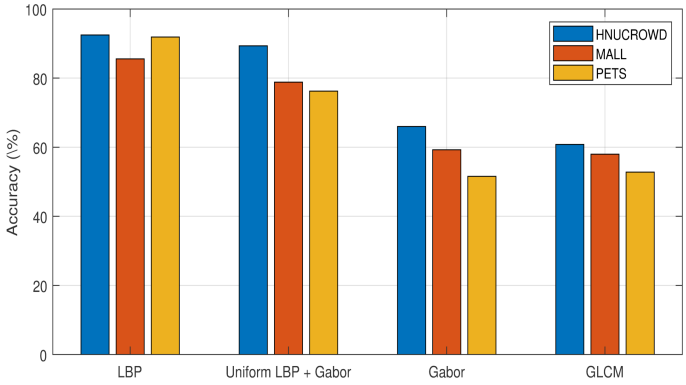

Fig. 3. Comparisons of different hand-crafted features (LBP, Uniform LBP + Gabor [12], Gabor, and GLCM) using RBF kernel SVM on the three datasets: HNUCROWD, MALL and PETS.

features, see Table III. As shown in the table, LBP features achieve good classification accuracy on the three datasets, with better performance on HNUCROWD dataset since the crowd in this dataset is more uniformally distributed in the scene than the other datasets. Using RBF kernel, the results are better compared to linear kernel, which complies with other previous works [12], [18]. Thus, SVM using RBF kernel is selected for the next experiments. To demonstrate the effectiveness of LBP as hand-crafted feature extractor for crowd density estimation, we compare LBP with other frequently used texture features: Gabor and GLCM. Also a combination between Gabor and Uniform LBP proposed in [12] is evaluated, see Fig.3. As depicted in the figure, it is clearly shown that LBP outperforms the other descriptors namely, GLCM, Gabor, and Uniform LBP + Gabor [12], which justifies the relevance of using LBP as hand-crafted features for crowd density classification.

C. *Results of learned features and analysis*

At this stage, we intend to evaluate the performance of learned features. These features are extracted from the second fully connected layer in all models and trained using SVM classifier. Precisely, the performance of the proposed CrowdCNN architecture and the pre-trained models on the different datasets are compared using SVM on both linear and RBF kernels, the results are reported in Table IV. From the obtained results, it is shown that the proposed CrowdCNN architecture slightly outperforms the three pre-trained models. The classification accuracy reaches 96.00%, 91.99%, and 95.94%, on HNUCROWD, MALL and PETS datasets, respectively. Also, the obtained results by pre-trained models

Dataset	Classifier	CrowdCNN	AlexNet	VGG-19	Inception-v3
HNUCROWD	Linear SVM	**96.00**	**95.17**	94.5	93.00
	RBF SVM	95.83	94.17	**94.83**	**94.67**
MALL	Linear SVM	91.03	88.78	**92.31**	**88.14**
	RBF SVM	**91.99**	**90.06**	90.06	**88.14**
PETS	Linear SVM	**95.94**	91.87	95.31	**94.38**
	RBF SVM	95.63	**93.44**	**96.25**	92.81

TABLE IV
THE CLASSIFICATION ACCURACY USING LEARNED FEATURES FROM THE PROPOSED CROWDCNN ARCHITECTURE COMPARED TO THREE DIFFERENT PRE-TRAINED MODELS (ALEXNET, VGG-19 AND INCEPTION-V3) BY EMPLOYING LINEAR AND RBF KERNELS OF SVM ON THE THREE EXPERIMENTED DATASETS.

are good, mainly the results of VGG-19 model. In overall, the results of learned features are better than the hand-crafted features with a significant margin which complies with our proposal and with the general trend in computer vision field.

Moreover, we include a comparison between learned features using different classifiers, namely, decision tree [39], Bagging predictors [40], KNN [41] and subspace KNN [42] see Fig. 4. Using different classifiers, it has been again demonstrated that the proposed CrowdCNN architecture outperforms the three pre-trained models, which corresponds to the previous results. Better performance is almost noticed using subspace KNN compared to the other classifiers. The results of this classifier on the three experimented datasets are close to those obtained using SVM classifier.

D. Discussion and evaluation of generalization

To summarize, extensive tests on different datasets and using various feature extractors are performed. The quantitative evaluation of the obtained results demonstrates the effectiveness of learned features (by means of a proposed CrowdCNN architecture and pre-trained models) for crowd level classification with a significant margin compared to LBP as hand-crafted features. Nevertheless, it has been shown that LBP achieves good performance compared to other texture features, which justifies the choice of this descriptor as hand-crafted features. In addition, by comparing learned features, better performance is almost noticed using the proposed architecture regarding the pre-trained models, except for VGG-19 model that performs equally well with the proposed architecture.

To better highlight the generic aspect of learned features vs. hand-crafted features, we evaluate the results by mixing the three datasets using SVM classifier (with RBF kernel). By doing that, high performance is achieved using the three pre-trained models, precisely, the obtained accuracies are 93.26%, 94.07% and 93.25% using AlexNet, VGG-19 and Inception-v3, respectively. The results of pre-trained models are high and even exceed the average results of the three datasets. Also, the proposed architecture achieves 93.90% as accuracy, which is a quite satisfactory result. Whereas, the performance of LBP as hand-crafted features is decreased to 87.26% only.

To conclude, as demonstrated from the obtained results, by mixing the three datasets, some results are affected more than others. Precisely, the fact of using heterogeneous data does not affect the performance of the pre-trained models. These results are expected since such models are trained on huge datasets and are capable of effectively encoding image representations in different situations and tasks. A limited impact of heterogeneous dataset is noticed on the proposed architecture as well thanks to the effectiveness of deep representations.

IV. CONCLUSION

In this paper, we focused on the problem of feature extractors to characterize the crowd texture. In particular, we employed deep learning models for automatic feature extraction in the crowd. To achieve this goal, we assessed different learned features compared to the commonly used hand-crafted features. Furthermore, we included a large comparative study between different classifiers to better prove the obtained results. The experimental results highlight a high performance of learned features (from the proposed architecture and pre-trained models) compared to LBP descriptor as hand-crafted features. In addition, it has been demonstrated from the obtained results using heterogeneous dataset that CNN features are effective enough for scene generalization. Hence, the representative power and the generalization ability of CNN features compared to hand-crafted features have been proven through this current paper for crowd density application, which complies with previous works for other classification tasks [21]–[23] in the literature.

There are several possible extensions of this research work. Since CNN features obtained from intermediate layers encode useful information, one potential perspective of this paper could be the fusion and the selection of multi-layer features for complementary aspect in order to achieve better performance. Also, a combination of hand-crafted and deep learning features can be studied as recently proposed in other other fields of application [43] [44]. Finally, this study can be practically useful for other video surveillance applications since the choice of features to describe an image content is a crucial component in any visual recognition task.

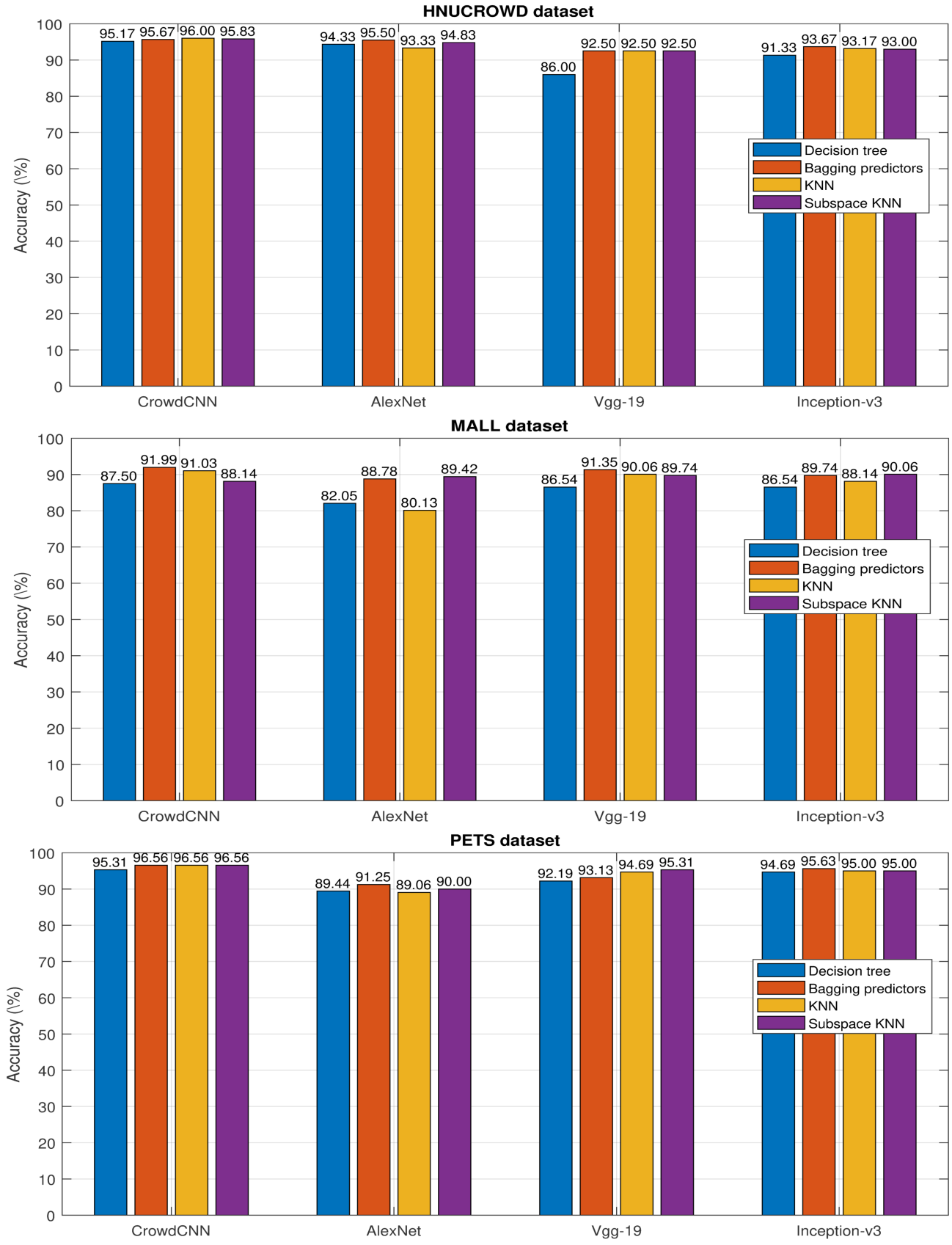

Fig. 4. Comparisons of the proposed CrowdCNN architecture with pre-trained models on different datasets (HNUCROWD, Mall, PETS) using different classifiers, namely decision tree [39], Bagging predictors [40], KNN [41], and subspace KNN [42].

REFERENCES

[1] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, March 2015.

[2] M. S. Zitouni, H. Bhaskar, J. Dias, and M. Al-Mualla, "Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques," *Neurocomputing*, vol. 186, pp. 139 – 159, 2016.

[3] H. Fradi, B. Luvison, and Q. C. Pham, "Crowd behavior analysis using local mid-level visual descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 589–602, March 2017.

[4] H. Fradi and J. Dugelay, "Towards crowd density-aware video surveillance applications," *Information Fusion*, vol. 24, pp. 3–15, 2015.

[5] F. Xiong, X. Shi, and D.-Y. Yeung, "Spatiotemporal modeling for crowd counting in videos," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5161–5169, 2017.

[6] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, vol. 107, pp. 3 – 16, 2018, video Surveillance-oriented Biometrics.

[7] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, Sept 2012, pp. 470–475.

[8] A. Polus, J. L. Schofer, and A. Ushpiz, "Pedestrian flow and level of service," *Journal of Transportation. Engineering*, vol. 109, pp. 46–56, 1983.

[9] K. Keung, L. Y. Xu, and X. Wu, "Crowd density estimation using texture analysis and learning," *IEEE International Conference on Robotics and Biometrics*, pp. 214–219, 2006.

[10] W. Ma, L. Huang, and C. Liu, "Crowd density analysis using co-occurrence texture features," in *5th International Conference on Computer Sciences and Convergence Information Technology*, Nov 2010, pp. 170–175.

[11] B. Zhou, B. Song, M. M. Hassan, and A. Alamri, "Multilinear rank support tensor machine for crowd density estimation," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 382 – 392, 2018.

[12] A. K. Pai, A. K. Karunakar, and U. Raghavendra, "A novel crowd density estimation technique using local binary pattern and gabor features," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug 2017, pp. 1–6.

[13] B. Zhou, F. Zhang, and L. Peng, "Compact representation for dynamic texture video coding using tensor method," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 280–288, Feb 2013.

[14] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[15] W. Ma, L. Huang, and C. Liu, "Advanced local binary pattern descriptors for crowd estimation," *Computational Intelligence and Industrial Application*, vol. 2, pp. 958–962, 2008.

[16] H. Yang, H. Su, S. Zheng, S. Wei, and Y. Fan, "The large-scale crowd density estimation based on sparse spatiotemporal local binary pattern," *IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2011.

[17] S. M. Mousavi, S. O. Shahdi, and S. A. R. Abu-Bakar, "Crowd estimation using histogram model classification based on improved uniform local binary pattern," *International Journal of Computer and Electrical Engineering*, vol. 4, pp. 256–259, 2012.

[18] H. Fradi and J. Dugelay, "A new multiclass svm algorithm and its application to crowd density analysis using lbp features," in *2013 IEEE International Conference on Image Processing*, Sept 2013, pp. 4554–4558.

[19] Z. Wang, H. Liu, Y. Qian, and T. Xu, "Crowd density estimation based on local binary pattern co-occurrence matrix," *IEEE International Conference on Multimedia and Expo Workshops*, 2012.

[20] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85 – 117, 2015.

[21] L. Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition*, vol. 71, pp. 158 – 172, 2017.

[22] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 1263–1266.

[23] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features for pedestrian, face and edge detection," *ICCV*, vol. abs/1504.07339, 2015.

[24] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 589–597.

[25] W. Weng and D. Lin, "Crowd density estimation based on a modified multicolumn convolutional neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–7.

[26] F. Xiong, X. Shi, and D. Yeung, "Spatiotemporal modeling for crowd counting in videos," *ICCV*, vol. abs/1707.07890, 2017.

[27] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *CVPR*, 2018.

[28] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 81 – 88, 2015.

[29] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 833–841.

[30] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 580–587.

[32] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1717–1724.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *in Proc. Int. Conf. Learn. Representations,*, vol. abs/1409.1556, 2015.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[36] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Dec 2009, pp. 1–6.

[37] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *BMVC*, 2012.

[38] B. Zhou, F. Zhang, and L. Peng, "Higher-order svd analysis for crowd density estimation," *Computer Vision and Image Understanding*, vol. 116, no. 9, pp. 1014 – 1021, 2012.

[39] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, Apr 2013.

[40] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[41] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theor.*, vol. 13, no. 1, pp. 21–27, Sep. 2006.

[42] T. K. Ho, "Nearest neighbors in random subspaces," in *Lecture Notes in Computer Science: Advances in Pattern Recognition*. Springer, 1998, pp. 640–648.

[43] S. Wu, Y. Chen, X. Li, A. Wu, J. You, and W. Zheng, "An enhanced deep feature representation for person re-identification," *WACV*, vol. abs/1604.07807, 2016.

[44] D. Yadav, N. Kohli, A. Agarwal, M. Vatsa, R. Singh, and A. Noore, "Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection," in *CVPR Workshops*, 2018.