Transfer learning for the classification of video-recorded crowd movements

Mounir Bendali-Braham, Jonathan Weber, Germain Forestier, Lhassane Idoumghar and Pierre-Alain Muller
IRIMAS, Université de Haute-Alsace, Mulhouse, France
Email: {first-name.last-name@uha.fr}

***Abstract*—The automatic recognition of a crowd movement captured by a CCTV camera can be of considerable help to security forces whose mission is to ensure the safety of people on the public area. In this context, we propose to fine-tune a model from the TwoStream Inflated 3D architecture, pre-trained on the ImageNet and the Kinetics source datasets, to classify video sequences of crowd movements from the Crowd-11 target dataset. The evaluation of our model demonstrates its superiority over the state-of-the-art in terms of classification accuracy.**

***Index Terms*—Video-surveillance, Crowd Behavior Analysis, Convolutional Neural Networks, Transfer Learning.**

I. Introduction

Either a culmination of a social protest or a cultural event, or an inevitable consequence of densely populated cities, crowd movements occur more and more in the public area [1]. The high frequency of these movements pushes the security forces to gain more control on them [1], [2]. Recent events have demonstrated the dangers of an uncontrolled crowd movement: a mismanaged crowd event can lead to heavy casualties [1].

In order to manage crowd movements, security forces can rely on the use of video-surveillance cameras [2]–[4]. The disposal of these cameras should cover a large part of the public area [5]. Although one of their most common uses is the acquisition of images that demonstrate criminal activity and their subsequent use for forensic purposes, the use that is beginning to be made of them is crowd analysis to predict abnormal situations [4]. However, despite the abundance of raw data from video-surveillance cameras, there is no unified model which can be used in all case-scenarios of crowd movements. This is due to the paucity of publicly available annotated datasets [6].

Today, due to its multiple successes, deep learning is trending in computer vision [7]. Although these methods appeared more than two decades ago, they are more and more used since the multiplication of their successes in image classification. One of the first and most notable success was realized by AlexNet [8], which achieved considerable performance in image classification when trained on the ImageNet dataset [9]. Although a part of computer vision, crowd behavior analysis did not benefit from the popularity of deep learning methods in computer vision. The scarcity of data and the lawful difficulty of obtaining them are one of the causes of this delay.

Recently, a team from the CEA (The French Alternative Energies and Atomic Energy Commission) has created a dataset called Crowd-11 [10]. This dataset, of over 6,000 video clips, is a major contribution to crowd behavior analysis, because it describes ten observable crowd movements in the public area or in large enclosed spaces such as airports or supermarkets. Successfully developing a statistical model capable of classifying these movements can be of great help for the security forces.

In this paper, we applied transfer learning to classify video sequences of crowd movements. We fine-tuned a model from the TwoStream Inflated 3D ConvNet (I3D) architecture [6] that had already been pre-trained on the ImageNet [9] dataset and the Kinetics [11] action recognition dataset, on what has been recovered from the Crowd-11 dataset. The fine-tuned TwoStream-I3D model is compared to a model from the 3D Convolutional Networks (C3D) architecture [12], which was pre-trained on the Sports-1m dataset and then fine-tuned on the same dataset. The rest of this paper is organized as follows: in Section II-A, we discuss the related work topics covered in crowd analysis. Afterthat, we present the Crowd-11 dataset in Section II-B, and then show the difference between the original dataset and what we could retrieve from it. We introduce transfer learning for video classification, in Section III, and we present the architectures for which we applied it. In Section IV, we explain the different experiments we undertook on Crowd-11 through k-fold cross validation and discuss the evaluation results.

II. Background

A. Related work

Crowd analysis has been part of computer vision research for more than two decades. Work in this area is divided into two broad categories: crowd statistics and crowd behavior analysis [13]–[15]

Crowd statistics:

- *Crowd counting:* This subtopic of crowd statistics consists of counting the number of individuals contained within a crowd in a scene [16].
- *Crowd density estimation:* Estimating crowd density in a scene can be of considerable help for crowd management [17].

Crowd behavior analysis:

- *Trajectories analysis:* This theme is part of what is mostly done in crowd behavior analysis [18]. Trajectories

 DOI: https://doi.org/10.1109/ISPA.2019.8868704

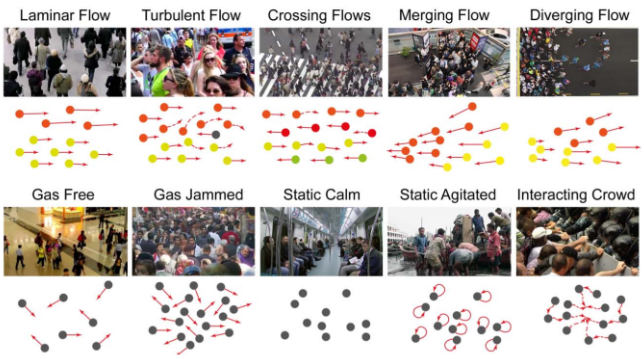

Fig. 1. Figure obtained from Dupont et al.'s paper illustrating the behavior of each crowd constituting a class in the Crowd-11 dataset [10].

analysis can be used for group detection [19], anomalous trajectories detection [20], and future trajectories prediction [21].

- *Group detection and behavior analysis:* After detecting groups, some works focus on the recognition of group actions [22]. Group detection and group behavior analysis are part of the mesoscopic approaches of crowd analysis, because a group is halfway between an individual and a crowd [23].
- *Anomaly Detection:* Anomaly detection can be done for any task of crowd analysis [24]. However, as Thida et al. specify, researchers did not agree on a unanimous definition of normality [25], because an anomaly in crowd analysis can range from recognizing abnormal events such as fights, traffic accidents, forgetting a luggage when leaving a train station, or witnessing a unusual event such as a pedestrian walking in the middle of a street.

Crowd analysis can rely on the manual extraction of visual cues. Most of this extraction is discussed in several reviews [13]–[15], [26]. The extraction of visual cues can refer to the computation of optical flow in a video clip, or contours detection, the detection of points/regions of interest in a single frame that can lead to pedestrian detection. Following this, an extraction of the different trajectories in a scene. More recently, this task, often subject to a number of omissions, has started being delegated to deep neural networks, because they are often able to spot significant visual cues better than hand-crafted methods [27].

B. The Crowd-11 dataset

Created by the CEA-LIST team [10], this fully annotated dataset contains more than 6,000 video clips. Video clips have variable resolutions ranging from 220×400 to 700×1250, and are based on a multitude of pre-existing sources. The videos are classified in 11 categories illustrated in the figure 1.

In what follows, we describe the behaviors corresponding to the 11 classes contained in the Crowd-11 dataset:

0) **Gas Free :** Individuals walking in all directions without encountering obstacles.
1) **Gas Jammed:** Congested Crowd.
2) **Laminar Flow:** Crowd walking in one direction.
3) **Turbulent Flow:** Crowd walking in a single direction and disturbed by an individual crossing the crowd in the opposed direction.
4) **Crossing Flows:** Two crowds crossing each other.
5) **Merging Flows:** Two converging crowds.
6) **Diverging Flow:** A crowd that splits into two crowds.
7) **Static Calm:** A crowd of static and calm individuals.
8) **Static Agitated:** A crowd of static, but agitated individuals.
9) **Interacting Crowd:** Two opposed crowds. This class contains violent scenes.
10) **No Crowd:** No human presence in the scene.

The videos originate mainly from three video hosting websites which are Youtube, Pond5[1], and GettyImages[2]. The rest comes from the following datasets: UMN SocialForce, AgoraSet, PETS-2009, Violent-Flows, Hockey Fights and Movies, WWW Crowd, CUHK Crowd, and Shanghai WorldExpo'10 Crowd. Most of these datasets are publicly available and easily accessible. However, we could not get videos from WWW Crowd, CUHK Crowd, and Shanghai WorldExpo'10 Crowd. Because of this, we were unable to retrieve the Crowd-11 dataset in its entirety. We could obtain approximately 90% of the original dataset. The distribution of the retrieved clips for each class, displayed in the comparative table I, shows that we did not endure a major loss of videos from the original dataset.

Label	Class name	#clips (original)	#clips obtained
0	Gas Free	529	477
1	Gas Jammed	520	508
2	Laminar Flow	1304	1189
3	Turbulent Flow	892	862
4	Crossing Flows	763	717
5	Merging Flow	295	267
6	Diverging Flow	184	189
7	Static Calm	737	686
8	Static Agitated	410	351
9	Interacting Crowd	248	153
10	No Crowd	390	370

TABLE I
COMPARISON BETWEEN THE NUMBER OF CLIPS PER-CLASS OF ORIGINAL CROWD-11 AND OURS.

III. TRANSFER LEARNING

Most of the time, transfer learning for the classification of video clips has been applied for action recognition in individual scenes [6], [12]. In this situation, the purpose is to transfer the knowledge learned from a source dataset to a target dataset belonging to the same topic. Dupont et al. [10] applied this operation by transferring the features that a model learned from an action recognition source dataset to a target dataset of crowd movements. The purpose of transfer learning is to transmit the features learned by a model from a source dataset to a target dataset [28].

[1]Pond5: https://www.pond5.com/
[2]GettyImages: https://www.gettyimages.com/

A. Implemented architectures

We selected three models to fine-tune from two architectures: C3D and TwoStream-I3D. The choice of the TwoStream-I3D architecture is mainly motivated by the good results that its models obtain compared to the C3D models when they perform action recognition in individual scenes on the UCF-101 and HMDB-51 datasets [6]. As the CEA team obtained their best results with the C3D architecture, its choice in our experiments is natural since we were not able to retrieve the Crowd-11 data set in its entirety. A pre-trained C3D model on Sports-1m got its best results when classifying Crowd-11 videos [10]. Therefore, this model represents for us the baseline result to improve during our experiments.

1) 3D Convolutional Neural Network: We decided to re-implement a version of the 3D Convolutional Neural Networks that correspond to the architecture described in [12]. The C3D architecture consists of 5 3D Convolutional layers. Each these layers is followed by a three-dimensional max pooling layer. These 5 first layers are then followed by 3 fully connected layers. The last layer has a softmax classification output made up of 11 classes.

As we have already mentioned, the CEA team gets their best performance with C3D after pre-training the model on Sports-1m [29]. The Sports-1m dataset is a dataset that contains 1 million videos from Youtube classified in 487 categories. Each category contains approximately 1,000 to 3,000 videos per class.

2) Two-Stream Inflated 3D Neural Network: Carreira and Zisserman propose the Two-Stream Inflated 3D ConvNets architecture [6]. This architecture was used to learn action recognition in individual scenes, where it obtained very good results compared to C3D. We use it to learn crowd movements recognition.

Carreira and Zisserman pre-trained a TwoStream-I3D model on Kinetics [11] and ImageNet [9]. By testing this model on the UCF-101 and HMDB-51 datasets, they significantly outperformed the performance of the pre-trained C3D models on Sports-1m [6]. In our situation, we decided to transfer the learned features of an RGB stream of the I3D architecture on the Kinetics and ImageNet source datasets to the Crowd-11 target dataset. We did the same for the TwoStream-I3D model by transferring the learned features of the RGB stream and the optical flow stream of the architecture to the target dataset. We extracted the optical flow of each video clip using the TV-L1 algorithm [30]. The architecture from which we derive I3D and TwoStream-I3D models is illustrated in Figure 2. It consists of two layers of three-dimensional (3D) convolutional layers supplemented by batch normalization layers, each of which is followed by a 3D max pooling layer. This bedrock is followed by a series of nine Inception modules whose internal characteristics vary slightly from one module to another. At the end, the output of the last Inception module is passed through a 3D average pooling layer, before going through a softmax output layer for the classification into 11 classes.

IV. Experiments on Crowd-11

In the experiments that we undertook, we decided for each architecture to fine-tune a pre-trained model and to train a model from scratch on Crowd-11. In the case of the C3D pretrained model, the pretraining was performed on the Sports-1m dataset. In the case of the I3D streams, the pretraining was performed on ImageNet and then on respectively the RGB version of Kinetics for the RGB stream and the optical flow version of Kinetics for the optical flow stream. Inspired by the training setting found on Tran et al. and Carreira et al. for respectively the C3D and the TwoStream-I3D models [6], [12], we chose the Stochastic Gradient Descent (SGD) as an optimization function, and fixed the learning rate (LR) to 0.003. The chosen loss function for these experiments is the categorical cross-entropy. In order to be very close to the training setup of C3D when trained from scratch or fine-tuned on Crowd-11 by Dupont et al. [10], we reproduced the LR gradual decrease by dividing it by 10 each 4 epochs. However, we did not follow this same policy for I3D and TwoStream-I3D. We chose to decrease the LR by 10 when the loss on the validation set did not improve. During the training phase, the number of epochs was fixed to 40 for C3D models, and 30 for the others, so as to maximize the opportunity of C3D models to get better scores. A model is produced at the end of each epoch. At the end of the training phase, we chose to keep the model that minimizes the loss function at the validation phase. When we applied fine-tuning, we did not freeze any layer of our models. We decided to avoid doing so, because the source datasets on which our models were pre-trained on differ a lot from the target dataset we intended to learn. Consequently, we were moved by the idea to backpropagate the training updates on all the weights of the networks we train. Contrary to Dupont et al. we did not apply data augmentation to train any of these models. Knowing that data augmentation is a regularization method, we wanted to observe to what extent our models could overfit the dataset [31]. Furthermore, we wanted to determine which classes could undermine the learning ability of our models, without reducing this issue using data augmentation. As we intend to use all the possible ways to augment our video data, we leave this question to a future work.

A. 5-fold cross validation :

Our version of Crowd-11 is made up of 1641 scenes. These scenes are split into 5769 video clips. To avoid scenes overlapping between folds, we kept all the clips from the same scene in the same fold. When we select a scene to add to any fold, our selection maintains a quantity of clips per class that is proportionately similar between all the folds with respect to the original quantities displayed in Table I. To train or fine-tune our models, we applied the 5-fold stratified cross validation. We divided the dataset into 5 proportionate similar folds in terms of the contained classes. For each iteration of cross validation, we chose 3 folds to form the training set, one for the validation set and a last one for the test set. At each iteration of cross validation, the test set changes. The validation set is chosen randomly among the 4 remaining folds. As we applied

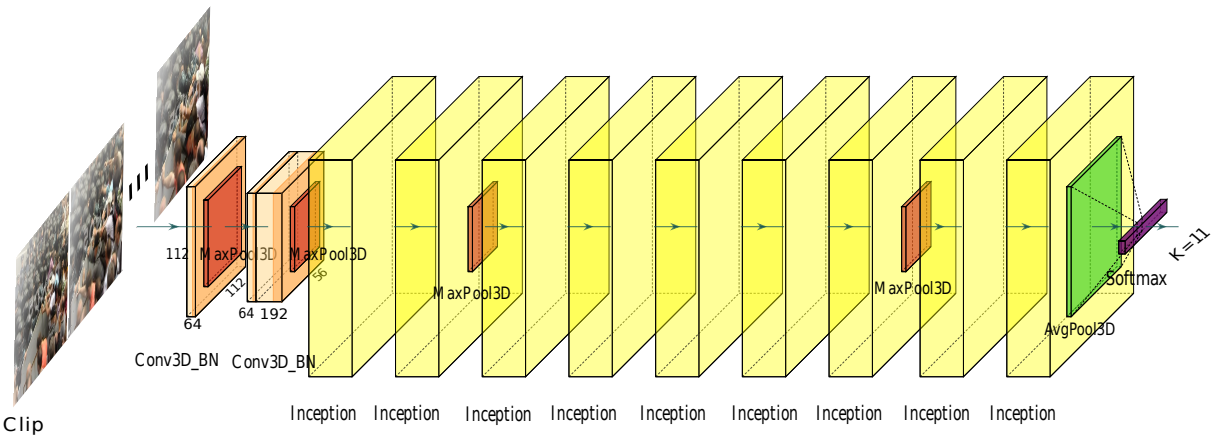

Fig. 2. Illustration of the Inflated 3D architecture. **Conv3D_BN** refers to the 3D convolutional layer followed by a batch normalization layer. **Inception** refers to an Inception module. **AvgPool3D** is a 3D average pooling layer.

Model	Training condition	Accuracy
C3D ours	Scratch	31.9%
C3D Dupont et al.	Scratch	46.9%
C3D ours	Pretrained	58.4%
C3D Dupont et al.	Pretrained	61.6%

TABLE II
COMPARISON BETWEEN OUR C3D AND DUPONT ET AL. [10]

Architecture	Training	Mean	Min	Max
I3D	Scratch	47.4%	40.1%	54.5%
C3D	Scratch	31.9%	28.9%	36.2%
TwoStream_I3D	Scratch	47.8%	44%	52.5%
I3D	Pretrained	59%	56.7%	60.1%
C3D	Pretrained	58.4%	57.6%	60.1%
TwoStream_I3D	Pretrained	**68%**	**66.2%**	**70.6%**

TABLE III
ACCURACY FOR 5-FOLD CROSS VALIDATION.

5-fold cross validation for each of our three models during the two prior training conditions : training from scratch, fine-tuning on top of a pre-trained model; we went through 30 training phases[3].

B. Discussion of the obtained results

From the boxplots illustrated in Figure 4, the variability of the models trained from scratch on Crowd-11 are less stable than those which were fine-tuned on the same dataset. According to the results displayed on Table II, we observe that C3D trained from scratch on Crowd-11 does not perform as well as Dupont et al.'s trained model. This may be due to the lack of information we have on the training setup they used to train their model, the slight difference between our two datasets, and the fact that we do not use video data augmentation. According to the results displayed on Table III, we find that the C3D and I3D models obtain almost the same results when classifying the video clips of the test set. C3D is exceeded with a margin of 0.6% by the I3D model. This slight difference in performance can be explained by the fact that the C3D architecture has 78 million parameters to train while the I3D architecture has 12 million parameters as well as a deep architecture. Moreover, we observe that the TwoStream-I3D model leverages favorably the use of optical flow in the fine-tuning phase. This is not totally demonstrated when it is trained from scratch. Compared to other models, TwoStream-I3D obtains the best results. From the confusion matrices displayed on Figure 3, we observe that the overall accuracy of each model suffers from almost the same categories where their score is at its lowest. Those categories whose id_labels range from 3 to 6, are respectively the Turbulent Flow, the Crossing Flows, the Merging Flow, and the Diverging Flow. We observe that the clips belonging to those classes, including the Laminar Flow class, are frequently mixed up with each other. While the Laminar Flow class does not suffer a lot from this confusion because the crowd follows a unique direction, the multiple key transitions that are illustrated in the four other classes can confuse the classification function. For instance, we observe that the Merging Flow class is not confused with the Diverging Flow class which demonstrates that the classification function learns well how to differentiate between these two behaviors. However, both of these classes are frequently confused with the Crossing Flows. When a crowd crosses with an other one, both of merging and diverging behaviors are observed. Furthermore, while the Crossing Flows is illustrated by $\approx$ 850 clips, the Merging Flow and the Diverging Flow classes are illustrated by $\approx$ 200 video clips each (as illustrated in Table I). This situation can lead to two classes being encompassed by a more global one like the Crossing Flows class.

V. CONCLUSION AND PERSPECTIVES

In this work, we investigated the ability of the TwoStream-Inflated 3D to benefit from its pretraining on the Kinetics and the ImageNet datasets to classify crowd behaviors on the

[3]The source code of this project is available here : https://github.com/MounirB/Crowd-movements-classification

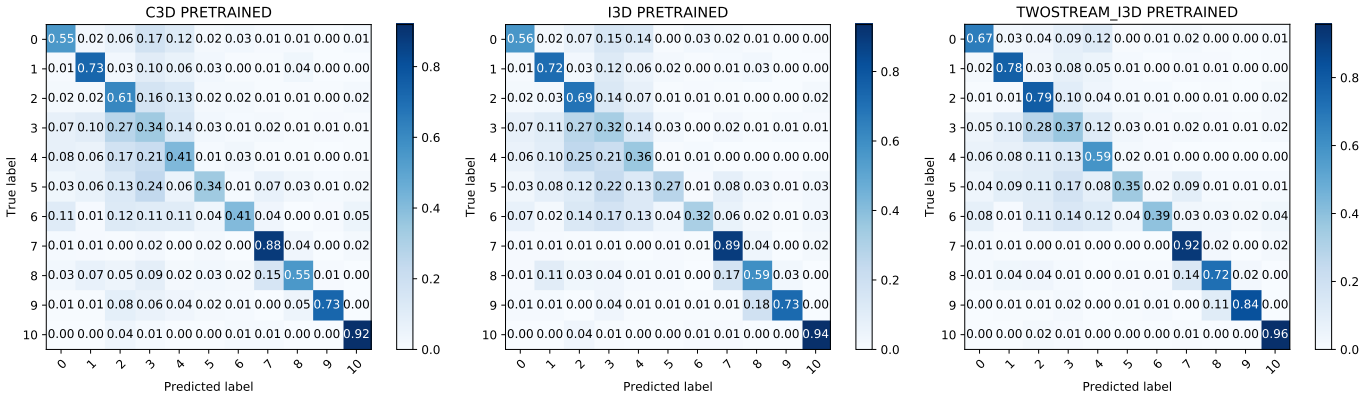

Fig. 3. Confusion matrices of the pretrained models computed following the 5-fold cross-validation.

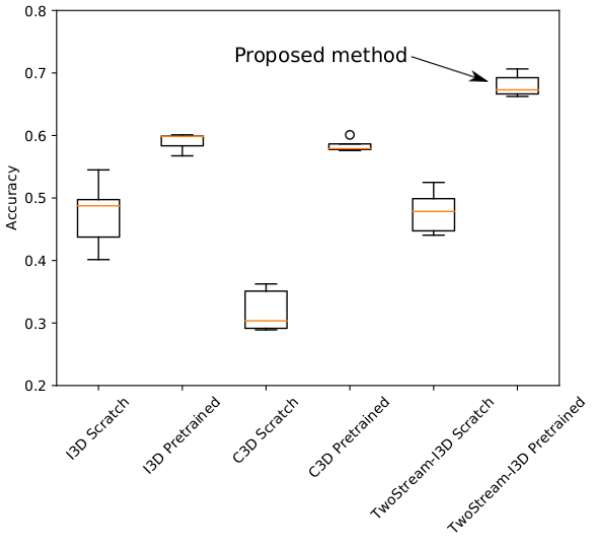

Fig. 4. Results obtained following the 5-fold cross validation step applied on Crowd-11 for each model.

Crowd-11 dataset. After transferring the weights learnt from its source datasets to the target dataset, the yielded model outperforms the state-of-the-art on Crowd-11 by a consequent margin of $\approx$ 10% accuracy. However, the obtained score cannot be considered as a precise decision tool for crowd management. On the basis of the results we have obtained, we intend to see to what extent we can improve them by testing the following methods:

- Applying video data augmentation;
- Augmenting the defective classes of the Crowd-11 dataset by adding video clips to them;
- Testing the models resulting from the Temporal 3D ConvNets (T3D) [32] and ActionVLAD [33] architectures, because the models from these architectures obtain scores exceeding 90% accuracy on the UCF-101 and HMDB-51 datasets;
- Modification of the Inflated 3D architecture via:
  - The addition of new Inception modules;
  - The hybridization of the I3D architecture with one of the two T3D or ActionVLAD architectures.
- Taking into account inputs from a preprocessing step, like the improved Dense Trajectories **(iDT)** [34], before proceeding to the training of models.

Acknowledgments

The authors would like to thank NVIDIA Corporation for the GPU Grant and the Mésocentre of Strasbourg for providing access to the GPU cluster. This work was supported by the ANR OPMoPS project (grant ANR-16-SEBM-0004) funded by the French National Research Agency.

References

[1] B. Krausz and C. Bauckhage, "Loveparade 2010: Automatic video analysis of a crowd disaster," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 307–319, 2012. https://doi.org/10.1016/j.cviu.2011.08.006.

[2] F. Porikli, F. Bremond, S. L. Dockstader, J. Ferryman, A. Hoogs, B. C. Lovell, S. Pankanti, B. Rinner, P. Tu, and P. L. Venetianer, "Video surveillance: past, present, and now the future [dsp forum]," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 190–198, 2013. https://doi.org/10.1109/MSP.2013.2241312.

[3] P. Drews, J. Quintas, J. Dias, M. Andersson, J. Nygårds, and J. Rydell, "Crowd behavior analysis under cameras network fusion using probabilistic methods," in *International Conference on Information Fusion*, pp. 1–8, 2010. https://doi.org/10.1109/ICIF.2010.5712106.

[4] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6479–6488, 2018. https://doi.org/10.1109/CVPR.2018.00678.

[5] J. Kritter, M. Brévilliers, J. Lepagnot, and L. Idoumghar, "On the optimal placement of cameras for surveillance and the underlying set cover problem," *Applied Soft Computing*, vol. 74, pp. 133 – 153, 2019. https://doi.org/10.1016/j.asoc.2018.10.025.

[6] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017. https://doi.org/10.1109/CVPR.2017.502.

[7] G. Seif, "Deep learning for image recognition: why it's challenging, where we've been, and what's next." https://towardsdatascience.com. Online; accessed 8-april-2019.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012. https://doi.org/10.1145/3065386.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

[10] C. Dupont, L. Tobias, and B. Luvison, "Crowd-11: A dataset for fine grained crowd behaviour analysis," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. 2017-July, (Honolulu, United States), pp. 2184–2191, 2017. https://doi.org/10.1109/CVPRW.2017.271.

[11] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *ArXiv*, 2017.

[12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015. https://doi.org/10.1109/ICCV.2015.510.

[13] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L. Q. Xu, "Crowd analysis: A survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, 2008. https://doi.org/10.1007/s00138-008-0132-4.

[14] S. Lamba and N. Nain, "Crowd monitoring and classification: a survey," in *Advances in Computer and Computational Sciences*, pp. 21–31, Springer, 2017. https://doi.org/10.1007/978-981-10-3770-2_3.

[15] J. M. Grant and P. J. Flynn, "Crowd scene understanding from video: a survey," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 2, p. 19, 2017. https://doi.org/10.1145/3052930.

[16] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 270–285, 2018. https://doi.org/10.1007/978-3-030-01234-2_17.

[17] X. Xu, D. Zhang, and H. Zheng, "Crowd density estimation of scenic spots based on multifeature ensemble learning," *Journal of Electrical and Computer Engineering*, vol. 2017, 2017. https://doi.org/10.1155/2017/2580860.

[18] W. Lu, X. Wei, W. Xing, and W. Liu, "Trajectory-based motion pattern analysis of crowds," *Neurocomputing*, vol. 247, pp. 213–223, 2017. https://doi.org/10.1016/j.neucom.2017.03.074.

[19] F. Solera, S. Calderara, and R. Cucchiara, "Socially constrained structural learning for groups detection in crowd," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 995–1008, 2015. https://doi.org/10.1109/TPAMI.2015.2470658.

[20] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 683–695, 2016. https://doi.org/10.1109/TCSVT.2016.2589859.

[21] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016. https://doi.org/10.1109/CVPR.2016.110.

[22] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971–1980, 2016. https://doi.org/10.1109/CVPR.2016.217.

[23] J. Shao, N. Dong, and Q. Zhao, "A real-time algorithm for small group detection in medium density crowds," *Pattern Recognition and Image Analysis*, vol. 28, no. 2, pp. 282–287, 2018. https://doi.org/10.1134/S1054661818020074.

[24] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS one*, vol. 13, no. 10, p. e0203668, 2018. https://doi.org/10.1371/journal.pone.0203668.

[25] M. Thida, Y. L. Yong, P. Climent-Pérez, H.-l. Eng, and P. Remagnino, "A literature review on video analytics of crowded scenes," in *Intelligent Multimedia Surveillance*, pp. 17–36, Springer, 2013. https://doi.org/10.1007/978-3-642-41512-8_2.

[26] 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, 2015. https://doi.org/10.1109/TCSVT.2014.2358029.

[28] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. https://doi.org/10.1109/TKDE.2009.191.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014. https://doi.org/10.1109/CVPR.2014.223.

[30] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Pattern Recognition*, pp. 214–223, 2007. https://doi.org/10.1007/978-3-540-74936-3_22.

[31] N. Dvornik, J. Mairal, and C. Schmid, "On the importance of visual context for data augmentation in scene understanding," *ArXiv*, 2018.

[32] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh, and L. Van Gool, "Temporal 3d convnets: New architecture and transfer learning for video classification," *ArXiv*, 2017.

[33] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 971–980, 2017. https://doi.org/10.1109/CVPR.2017.337.

[34] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558, 2013. https://doi.org/10.1109/ICCV.2013.441.

[27] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: a survey," *The Visual Computer*, pp. 1–24, 2018. https://doi.org/10.1007/s00371-018-1499-5.