Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Multilevel Data Enhancement Module For Image Quality Improvement in Group Activity Recognition

Yogita K. Desai[1]* and Dr. Amol D. Potgantwar[2]

[1] Department of Computer Engineering, MET BKC Institute of Engineering, Bhujbal Knowledge City, Adgaon, Nashik – 422003, Maharashtra, India, Savitribai Phule Pune University (SPPU), Maharashtra, India
ORCID: 0009-0007-8462-9293, Email: yogitadesaip23@gmail.com,
[2] Department of Computer Engineering, Sandip Institute of Engineering and Research Centre, Nashik – 422213, Maharashtra, India, Savitribai Phule Pune University (SPPU), Maharashtra, India, ORCID: 0000-0002-3621-4639, Email:amolp639@gmail.com
*Corresponding/Submitting Author
Yogita K. Desai
Department of Computer Engineering, MET BKC Institute of Engineering, Bhujbal Knowledge City, Adgaon, Nashik – 422003, Maharashtra, India, Maharashtra, India, Email: yogitadesaip23@gmail.com , keshavd.phdcomp_ioe@bkc.met.edu, ORCID: 0009-0007-8462-9293

Abstract
The goal of Group Activity Recognition (GAR) is to detect activities carried out by groups based on visual information. GAR is applied in many fields such as surveillance, crowd observation, sports analysis, and autonomous systems. Image quality has a significant impact on visual features. States of pre-processing methods are often treated only as a detail of practical implementation of the GAR system, depriving the pre-processing stage of a scientific definition. This paper presents a Multi-Level Data Enhancement Module (MDEM) defined as a compact and repeatable method of the pre-processing that consists of four stages: stochastic data augmentation, local contrast boosting, noise suppression, and channel-wise normalization. The proposed module is defined as a composite transformation performed at the pixel level. MDEM is tested on frames from the Collective Activity Dataset and Volleyball dataset with different characteristics including light contrast changes, camera movements, blurring, and image compression. The qualitative and quantitative analysis of the intermediate processing stages prove that better local contrast can be achieved while maintaining better uniform representation through noise suppression. The new method developed has a clear set of procedures for preprocessing that can be linked to different neural network approaches for GAR. Future research will concern quantitative ablation studies, temporal consistency in video Processing, and data-driven enhancement methods.

Keywords: Data Augmentation; Group Activity Recognition; Image Enhancement; Multi-Level Data Enhancement Module; Noise Reduction.

1. Introduction

Multiple people interact with each other in a group and this is captured with the help of cameras in the form of videos. What activity they are exactly per forming is the point of research in the field of computer vision. Many of the used areas for this are crowd monitoring, surveillance, autonomous driving, sports analysis etc. It differs from human activity recognition in that it not only captures the action of an individual person but every individual's interacting behaviour with the other persons in a group along with capturing of information over time and space dimensions For example, for sports a team of volleyball executing a “Spike” or a crowd “waiting” at some checkpoint. Visual appearance motion cues are important factors for GAR which can be extracted by Convolutional Neural Networks (CNN) backbone architectures. Two benchmarks have attracted attention of most of the researchers: Collective Activity Dataset (CAD) [1] introduced by Choi, Shahid and Savarse containing real world surveillance style footage of pedestrians with activities such as crossing, walking, talking, waiting and queueing. Another is sports dataset volleyball introduced by Ibrahim et al., [4] providing player level action annotations together with team level group activity labels drawn from volleyball footage. These datasets have quite different visual considerations, CAD sequences captured by hand-held and fixed

surveillance cameras under uncontrolled illumination while footage of volleyball initiated from broadcast video subject to compression artifacts, motion blur, and non uniform stadium lighting. It's important to address this heterogeneity as it may lead to capturing nuisance variation like lighting, jitter, noise rather than actually defining semantic cues for group activity recognition.

Learning of the network depends on the quality of the input frame to be given as input for CNN backbone network as prediction depends on visual appearance and motion cues. GAR research mainly focused on innovations in architecture from spatiotemporal models[1]-[3] to recurrent networks hierarchical in nature[4], graph based architecture[9],[10] and also transformer based architectures [11],[12] on mostly concentrated CAD and volleyball datasets. Preprocessing is applied to raw frames to backbone networks which leads to comparatively less attention. Data augmentation, noise reduction, contrast correction, and input standardization are individually well founded with broader literature of computer vision [5]–[7], [13], [16]. But for GAR related work they are typically applied in a specific manner, individual implementation, rather than reusable explicitly specified and data agnostic pipeline.

Motivated with this gap, Multi-Level Data Enhancement Module (MDEM) is introduced, a multi level compact preprocessing pipeline that consists of data augmentation, contrast enhancement, noise reduction and standardization into a single reproducible transformation chain. MDEM is backbone agnostic as it applies at pixel level on individual frames prior to feature extraction which can compose with any GAR model whether its recurrent, graph based, transformer based without any modification in model.

This research work contributes to threefold manner. First, MDEM formalization is done as a four stage composite operator $I_{enh} = S(N(E(A(I_t))))$ which provides an exact mathematical definition at each level. Second, MDEM is implemented and applied to sample frames from both CAD and Volleyball benchmark datasets. Which easily visually and statistically characterizes the effect at each level with the underlying pixel distribution. Third, this protocol is offered as a reproducible template for future work, directly addressing the absence of standardized, transferable preprocessing evaluation practice identified in Section 2.2, rather than as a substitute for the empirical results themselves.

The remainder of this paper is organized as section 2 reviews related work on GAR with benchmark datasets and practices for preprocessing. Section 3 describes two widely used datasets and also used in this study. Section 4 presents MDEM pipeline in detail. Section 5 presents implementation details. Section 6 discusses qualitative results and proposed quantitative evaluation protocol. Outline of future work and conclusion is discussed in section 7.

2. Related Work

2.1 Group Activity Recognition Datasets

Collective Activity Dataset (CAD) is featured as a benchmark for GAR. Spatio-temporal relationship along with CAD dataset is introduced by Choi et al. [1] using pose and trajectory context of surrounding individuals. Extension of this work was done by Choi and Savarese[2] for collective activity recognition and later proposed a framework for multi target tracking and collective activity recognition [3]. These early approaches rely on pose models and hand crafted trajectories.

Ibrahim et al. [4] drives a shift in end to end deep learning of GAR by proposing a two stage hierarchical deep temporal model with long short term memory (LSTM) unit[14]. First stage is person level LSTM which models individual actions and group level LSTM aggregates individual actions into group level actions. Beside this Ibrahim et al. presented the Volleyball Dataset as a large dataset which provides both person level and group level actions. CAD and volleyball dataset are the most popular datasets for group activity recognition.

The number of architectures are evaluated on both these benchmark datasets. Hierarchical relational network proposed by Ibrahim and Mori [9]. In GAR, activity depends on interactions among individuals so Wu et al. [10] proposed actor relation graphs (ARG) which learns pairwise relations in a graph. Actor transformer models [11] proposed by Gavrilyuk et al. where self attention is used to combine pose and appearance streams.Li et al. [12] further extended transformer-based spatio-temporal reasoning with clustered attention in GroupFormer, reporting state-of-the-art results.

More recent architectures have continued to advance the state of the art on both benchmarks. Ibrahim and Mori [9] proposed hierarchical relational networks for joint group-activity recognition and retrieval. Wu et al. [10] introduced actor relation graphs (ARG) that learn pairwise actor interactions through graph convolution rather than a hand-designed recurrent hierarchy. Gavrilyuk et al. [11] proposed actor-transformer models that

combine pose and appearance streams through self-attention. Li et al. [12] further extended transformer-based spatio-temporal reasoning with clustered attention in GroupFormer, reporting state-of-the-art results at the time of publication. This constant movement in the field of architecture demonstrates the ongoing general analysis of human behavior by Aggarwal and Ryoo [8], while CAD and Volleyball remain the standard measures in GAR research, which explains their utilization in this study.

2.2 Preprocessing and Data Augmentation in GAR

To improve the generalization of deep visual models data augmentation is widely used. A comprehensive survey of augmentation techniques with geometric transformations, jitter and mixed strategies are given by Shorten and Khoshgoftaar [6]. Choice of augmentation strategy depends on target task and domain which is largely absent for GAR task.

Methods for enhancing contrasts, such as Contrast-Limited Adaptive Histogram Equalization (CLAHE), which was described by Zuiderveld [5], are commonly used as an effective approach to restore local contrast in images that have variable illumination. Since CLAHE is a tile-based and clip-limited approach, it can be particularly useful for improving Generalized Autoregressive (GAR) video material where the lighting is usually heterogeneous, occasionally because there can be some local shadowing associated with a scene like the one from CAD or a stadium sport. However, this component is not mentioned in many research papers as a separate component of GAR video processing.

Classical noise-suppression filters, such as Gaussian smoothing and median filtering still hold their grounds to be cost-effective preprocessing techniques in the removal of noise from sensors and compression ahead of feature extraction, as shown by standard references in digital image processing such as Gonzalez and Woods. All these filters will be applicable to different sources of noise, which means that the use of these two filter families together can produce even greater results, although there are few examples of such use in GAR systems.

Ultimately, performing channel-wise normalization of input signals with the help of statistics computed on large collections of images — most commonly, mean and standard deviation obtained from the ImageNet dataset [7], [16] — has long been viewed as an effective means for achieving better stability of gradient-based optimizers utilized in Convolutional Neural Networks like ResNet [13], where batch normalization [15] guarantees the conformity of the input signal to the standard for normalization that has been developed in connection with training with the ImageNet data.

Even though all four techniques are quite well studied, the existing publications relevant to Agri-food-Gunda Analysis rarely present any of the aforementioned techniques as something more than the detail of the execution of the experiments.

Apart from the general-purpose preprocessing techniques mentioned earlier [5], [6], it is interesting to note what the GAR architecture papers presented in Section 2.1 talk about regarding their preprocessing. Table A below provides a summary on this point. Among the six GAR papers reviewed that cover a period of about a decade of architectural developments from Choi et al.'s initial manual CAD technique [1] to GroupFormer [12], there is none that describes a complete, well-justified preprocessing method as suggested in MDEM theory. Early approaches that rely on the use of hand-crafted features [1]–[3] did not perform any CNN-style preprocessing since these methods use pose and trajectory descriptions instead of original pixel information.

In addition to the previously mentioned preprocessing techniques for general use [5], [6], it might be of interest to see what has been mentioned in the GAR-related papers discussed in Section 2.1 regarding preprocessing. A summary of this information is presented in Table A below. Out of the six GAR-related papers published during the last decade that discuss the architectural progression starting from the first manual CAD method [1] developed by Choi et al. until GroupFormer [12], none of the papers provide a thorough and comprehensible method of preprocessing as defined by the MDEM theory. Early approaches depending on hand-prepared features [1]–[3] did not use any CNN-processed preprocessing techniques as they operated with the help of pose and trajectory rather than with original pixel data. Table 1 gives preprocessing methods by representative GAR studies.

Table 1. Preprocessing methods used or reported by representative GAR studies

Study (Ref.)	Year	Feature Backbone
Choi et al. — CAD [1]–[3]	2009–2012	None (hand-crafted STIP/pose/trajectory features)

Ibrahim et al. — Hierarchical Deep Temporal Model [4]	2016	Fine-tuned AlexNet (person-level CNN)
Ibrahim & Mori — Hierarchical Relational Networks [9]	2018	CNN person-level features (relational network)
Wu et al. — Actor Relation Graphs (ARG) [10]	2019	Inception-v3 / VGG-16 (ImageNet-pretrained), RoIAlign features
Gavrilyuk et al. — Actor-Transformer [11]	2020	I3D (Kinetics-pretrained) + 2D CNN; RGB and optical-flow streams
Li et al. — GroupFormer [12]	2021	Inception-v3 or I3D backbone, RoIAlign

Table 2 gives parameters used by the MDEM module.

Table 2 Parameters for MDEM module

MDEM : MultiLevel Data Enhancement Module	
Feature Backbone	Backbone-agnostic (pixel-level, applied ahead of any backbone)
Data Augmentation	**Explicit** — 6 stochastic transforms, p = 0.6, fully parameterized (Sec. 4.2)
Contrast Enhancement	**Explicit** — CLAHE on LAB-L channel, clip limit 2.0, 8×8 tiles (Sec. 4.3)
Noise Reduction	**Explicit** — Gaussian (3×3, σ=0.8) + median (3×3) cascade (Sec. 4.4)
Standardization / Normalization	**Explicit** — ImageNet mean/std, fully specified (Sec. 4.5)

Interpretation

The table clarifies what Section 2.2 claims qualitatively: no preceding GAR article outlines, dissects, or makes rationale for full and complete process of four stages of preprocessing (data augmentation + contrast improvement + noise suppression + data standardization) with such thoroughness as is done in MDEM article. The only aspect that is universally mentioned in the mentioned papers is backbone choice, which allows one to infer but not formulate the process of standardization according to ImageNet; nothing regards contrast improvement and noise suppression processes in all six papers' descriptions. This table can be presented as soon as Section 2.2 ends.

3. Datasets

Collective Activity Dataset (CAD) consists of 44 sequence videos with fixed and hand-held consumer cameras. As a portion of the sequences are hand held CAD frames exhibited with camera shake, variable framing and illumination between frames is inconsistent. Which makes it essential for evaluating robustness of a preprocessing pipeline.

On the other hand volleyball dataset with 4830 clips from 55 volleyball videos from broadcast televisions. As footage captured from broadcast television, frames are with fast camera pans, motion blur for rapid player movement etc. are there.

Table 2 gives details of these two (CAD and Volleyball) Datasets.

Table 2. CAD and Volleyball dataset details

Dataset	No. of Videos	Activities (Count)	Data Type	Key Features
CAD	44	5 (crossing, walking, waiting, talking, queueing)	Surveillance video clips	2481 clips, annotated frames with pose info
Volleyball	55	8 (Left/right set, left/right spike, left/right pass, left/right winpoint)	Sports video clips	Group & individual annotations for spatiotemporal activities

4. Proposed Methodology: MDEM

4.1 Overview and Formulation

Every raw frame I_t is processed by MDEM before its giving for feature extraction. MDEM is defined as a sequence of functions provided as input to the feature extraction step.

$$I_{enh} = \text{Standardize}(\text{NoiseReduce}(\text{Enhance}(\text{Augment}(I_t))))$$

With each stage being implemented as an independent, parameterized operator so that individual levels can be enabled, disabled without any alteration in the remaining pipeline. can be enabled, disabled, or re-tuned without altering the remaining pipeline. Order of level is Augmentation, Enhancement, noise reduction and standardization. Intentional order denotes that data augmentation is applied so that subsequent enhancement and noise reduction applies on the same geometric and photometric variations observed by model at training time. Before denoising enhancement is applied as local contrast can introduce or amplify frequency artifacts which will be removed at denoising level. Denoising precedes standardization so that final normalization statistics are computed on a clean signal in spite of noise corrupted pixel values. Fig.1 shows different MDEM levels with the output as an enhanced and standardized image.

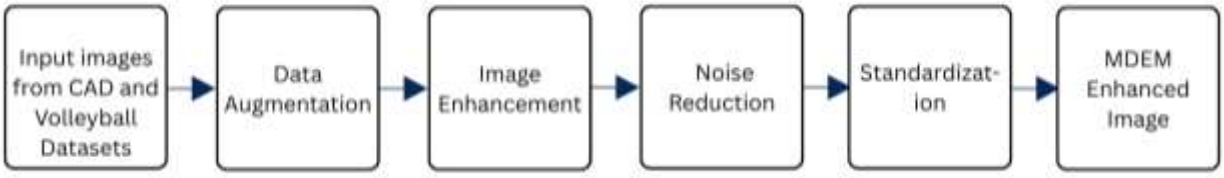

Fig.1 MDEM Architecture with different levels

4.2 Level 1: Data Augmentation

A is the Augmentation operator applies a subset of six different transformations, each triggered with probability p (p = 0.6), to expose a model to realistic nuisance variation during training

- Rotation (R) : The frame is rotated with an angle drawn uniformly from -12 degrees, +12 degrees, by reflective border padding so as to avoid artificial black borders.
- Horizontal flip F – the frame reverses along the vertical axis, demonstrating the left-right balance found in various collective and team-sport activities.
- Perspective warp P – all four corners of the image are jiggled freely by as much as 6% of the frame size, and the perspective transformation is fitted and applied herewith, giving a simulated viewpoint change which looks like shooting with a hand-held or panning camera.

- Scale S – the image amplifies by any factor sampled from [0.9, 1.1], then rescales and cuts to the original resolution by reflective padding or center cropping, thus mimicking differences in subjects' distance from the camera.
- Brightness B – the pixel values are multiplied by any factor sampled from [0.8, 1.2] and brought back within the range of possible value, thus giving an impression of changed lighting.
- Contrast C – the pixel values are standardized to the frame average value ones (by a factor sampled from [0.8, 1.2]), which gives the impression of change of the contrast due to differences in lighting and sensor performance.

For the sequence of levels in order {R, F, P, S, B, C}, the augmented image is $I_{aug} = A(I_t) = f_6(f_5(\dots(f_1(I_t))))$, where every f_i is applied with probability p. This produces a combinatorially large space of augmented appearances from a single source frame which also preserves the semantic contents for correct activity labeling.

4.3 level 2: Image Enhancement

The enhancement operator E carries out two things: it standardizes the resolution spatially and corrects the local contrast adaptively. First, the augmented image is resized into an 224 x 224 resolution image using bilinear interpolation so that it fits the input requirements of the convolutional backbone used. Next, the image once it has been resized is transformed from the BGR color space to the LAB color space, and CLAHE (Contrast-Limited Adaptive Histogram Equalization) is applied to the L channel only with a clip limit of 2.0 and a tile grid of 8 x 8 before converting the image back to the original BGR color space. Since the process takes place on the L channel alone, the a and b channel chrominance of the images is maintained and only local illumination and contrast changes are corrected, for example, the shadow, glare and lighting from the stadium are altered, while the information provided by colors remains the same in terms of team identification in the Volleyball Dataset.

4.4 Level 3: Noise Reduction

The N operator reduces noise in high-frequency sensors and compression noise through a classical filtering method. It starts with the application of a Gaussian filter on the image at stage two. The Gaussian filter is isotropic and has a kernel with a size of 3 x 3 and a standard deviation of 0.8. Thus, $I_g = G_{sigma} * I_{enhS2}$, where G_{sigma} is the kernel and * indicates convolution. After that, a median filter is applied to the Gaussian-smoothed picture; thus, $I_{nr} = \text{Median}_{(3X3)I_g}$. The combination of ordinary linear and non-linear filters is used so as to eliminate together two different kinds of noise sources, namely noise that comes from both sensor noise and impulse compression noise.

4.5 Level 4: Standardization

The last step transforms the cleaned-up and improved picture into a stable numerical tensor that is ready for modeling purposes. The pixel values are changed from the integer range of [0, 255] to the floating-point range of [0, 1]; the channel order goes from BGR to RGB; and the standardized value of each channel is brought to zero mean and unit variance based on constant per-channel statistics (with mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]), in full conformity with ImageNet-prior models:

$I_{enh} = (I_{nr} / 255 - \text{mean}) / \text{std})$

The final standardized tensor that is represented by I_{enh} in the pipeline is sent to the GAR feature extractor for further processes. Given the fact that the standardization statistics are in line with that of the usual use of pretrained models, images that have been processed by the MDEM are good for transfer learning purposes. Fig. 2 shows sample images processed by MDEM for CAD and Volleyball Dataset.

2(a) Sample image processed by MDEM from CAD Dataset

2(b) Sample image processed by MDEM from CAD Dataset

Fig. 2 Sample images Processed by MDEM

4.6 Algorithmic Summary

The following is a summary of complete MDEM produced by every level independently.

Input: Raw BGR color frame I_t, augmentation flag train

Step 1: if training is true, compute $I_{aug} = A(I_t)$ using stochastic operator sequence (section 4.2)

Step 2: Compute $I_{enh1} = E(I_{aug})$ via resize and CLAHE (section 4.3)

Step 3: Compute $I_{nr} = N$ (compute I_nr = $N(I_{enh1})$ with gaussian and median filtering (section 4.4)

Step 4: Compute $I_{enh} = S(I_{nr})$ with BGR to RGB conversion and channel wise standardization (section 4.5)

Output: Model ready tensor I_{enh} with intermediate representations I_{aug}, I_{enh1}, I_{nr} for diagnostic visualization.

5.Implementation Details

The code for MDEM was developed using the following tools: OpenCV allows to perform all operations related to image processing, NumPy helps to work with arrays and visualize them through Matplotlib. The footage used in the project comes from CAD and Volleyball datasets, both datasets were taken from local installations. Within each dataset, the first step involved finding the frames by looking through folders recursively within the range of feasible images (.jpg, .jpeg, .png, etc.). The seeds of the random libraries in Python were fixed (seed=42) making it possible to repeat the stochastic augmentation step independently of the run.In terms of qualitative analysis, all stages of processing (original, error-corrected, improved, reduced noise, standard for visuals) were run and compared on CAD and volleyball frames for testing. For batch processing, a set number of sampled images were processed through the pipeline, with three intermediate images output as JPEG and the final processed form as a NumPy array to facilitate further analysis later on. All parameters for enhancement, noise reduction, and standardization that were listed in Section 4 have been kept the same for both tests so that an ambivalent evaluation of the process can be made.

6. Results and Discussion

By inspecting the outputs of each MDEM stage on the CAD and Volleyball samples, the expected outcomes of each stage can be observed. The augmentation stage creates various controlled changes such as rotation, mirroring, perspective alteration, and brightness/contrast adjustments; however, the key elements of the overall scene remain unchanged and the results confirm that the parameter ranges given in Section 4.2 are safe and do not affect the activity-relevant information. The enhancement stage ensures proper contrast levels for the underexposed regions in the CAD and Volleyball sample frames without generating halos and noise that is usually generated due to global histogram equalization.The noise-reduction stage results in a visibly smoother picture that has lower levels of high-frequency noise and less visibility of blocky compression artifacts, which is most explicitly visible with low-resolution CAD frames, whereas crude geometrical boundaries (player outlines, court borders, drawing lines) are well preserved. Finally, if the tensor is rescaled to be displayed, we obtain the Gaussian distribution with mean close to zero for each pixel in each frame.

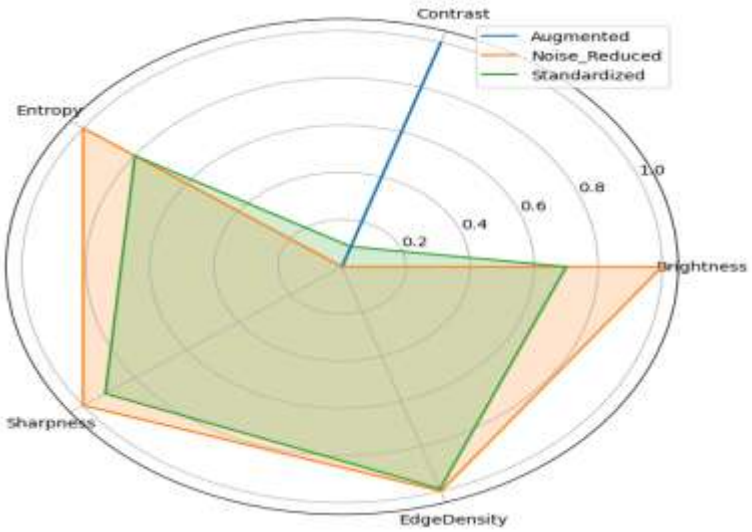

Fig. 3 Effect of preprocessing stages on image quality characteristic

The figure 3 shows how the processing stages of augmentation, noise reduction, and standardization influence the important features of the image. The Augmented images are characterized by high contrast as well as high brightness. In turn, the Noise-Reduced images contribute to the image sharpness and entropy levels, which confirms the retention of important structure details after noise removal. The Standardized images demonstrate very high edge density and entropy values while having balanced brightness and contrast. In summary, we can conclude that the implemented techniques are successful in the image quality improvement that is achieved thanks to eliminating unwanted noise while retaining the structural and texture features necessary for the analysis.

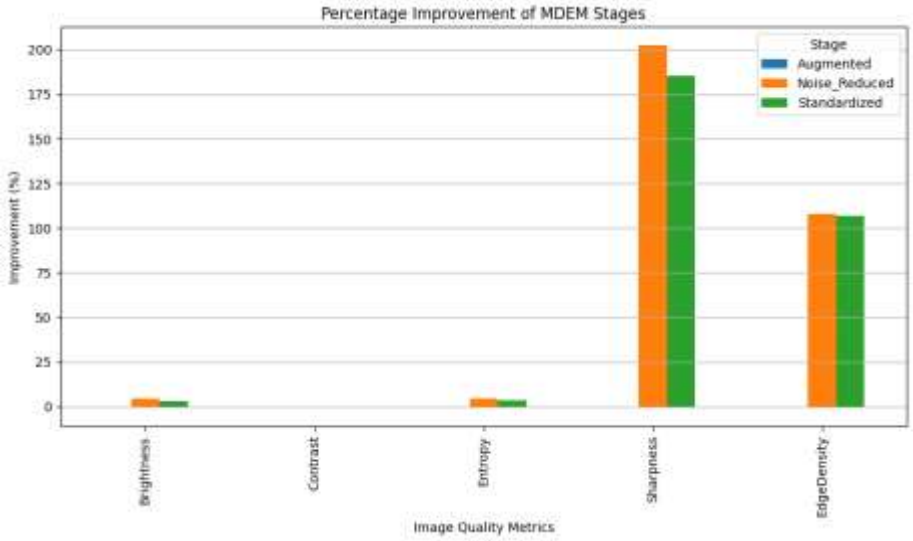

Fig. 4 Percentage Improvement in Image Quality Metrics Across MDEM Stages

The graph in fig. 4 signifies improvement in major image qualities at varied phases of implementation of the multiple domain enhancement module (MDEM). The findings suggest noise suppression and standardization greatly improve sharpness and edge density, with sharpness improved by nearly 203% and 185% and edge density by around 108% and 107%, respectively. In comparison to these parameters, brightness and entropy are insignificant, implying that the main gain of the MDEM preprocessing process lies in processing the structure and edge-related parameters rather than changing the overall intensity or data content. Minimal changes in contrast imply that the preprocessing phase keeps the original contrast features intact while

significantly enhancing the clarity and structural parameters of images. To sum up, the data received suggest that MDEM enables better quality of images and enhances the recognizable aspects of the input images.
The MDEM design intentionally separates different elements, the augmentation stage helps diversify training data which is good for generalization, the enhancement and noise reduction stages minimize other types of visual variations that can confuse classifiers, while standardization is responsible for the input statistics suitable for the optimizer. Since each processing component is specified separately without any tuning for a particular dataset this method should work well for other kinds of visual data, such as for new sports analytics or crowd surveillance data. However, the limitations of this design are that all processing stages work on very early representations, thus it would be reasonable to expand the approach to the temporal level, where the same transformations are applied across all frames of a video.
Additionally, it is worth mentioning that the enhancement and noise-reduction parameters selected for the current study (CLAHE clip limit 2.0, Gaussian sigma 0.8, and 3 x 3 kernel sizes) were based on the commonly used default parameters provided in the literature rather than made based on the dataset; automatic hyperparameter tuning, guided by the quantitative approach presented may help improve the results obtained with the methods in the future.

7. Conclusion and Future Work

In this paper, we introduced a Multi-level Data Enhancement Module (MDEM) which comprises four phases of preprocessing images of group activities and uses stochastic augmentation, LAB-space based adaptive contrast enhancement, Gaussian and median noise reducing, and the module of standardizing each channel combined into one clearly defined and generalizable mechanism. We have described all steps of the module clearly, obtained reproducibly, and applied it to the example pictures from the Collective Activity Dataset and Volleyball Dataset.According to the qualitative results, MDEM improves local contrast, reduces sensor and compression noise and creates a numerically stable ready-for-model tensor on both data sets, without any parameter tweaking per data set. Future plans include implementing the quantitative measurement protocol on several GAR backbones for identifying parts of every MDEM stage responsible for additional accuracy of the process

References

[1] Wongun Choi, K. Shahid and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 2009, pp. 1282-1289, doi: 10.1109/ICCVW.2009.5457461.
[2] Choi, K. Shahid and S. Savarese, "Learning context for collective activity recognition," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 3273-3280, doi: 10.1109/CVPR.2011.5995707.
[3] Choi, W., Savarese, S. (2012). A Unified Framework for Multi-target Tracking and Collective Activity Recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7575. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33765-9_16
[4] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat and G. Mori, "A Hierarchical Deep Temporal Model for Group Activity Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 1971-1980, doi: 10.1109/CVPR.2016.217.
[5] K. Zuiderveld, “Contrast Limited Adaptive Histogram Equalization,” in Graphics Gems, Academic Press, 1994, pp. 474–485. https://doi.org/10.1016/B978-0-12-336156-1.50061-6
[6] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), Vol. 1. Curran Associates Inc., Red Hook, NY, USA, 1097–1105. https://dl.acm.org/doi/10.5555/2999134.2999257
[8] J.K. Aggarwal and M.S. Ryoo. 2011. Human activity analysis: A review. ACM Comput. Surv. 43, 3, Article 16 (April 2011), 43 pages. https://doi.org/10.1145/1922649.1922653
[9] Ibrahim, M.S., Mori, G. (2018). Hierarchical Relational Networks for Group Activity Recognition and Retrieval. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds) Computer Vision – ECCV 2018. ECCV

2018. Lecture Notes in Computer Science(), vol 11207. Springer, Cham. https://doi.org/10.1007/978-3-030-01219-9_44
[10] J. Wu, L. Wang, L. Wang, J. Guo and G. Wu, "Learning Actor Relation Graphs for Group Activity Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 9956-9966, doi: 10.1109/CVPR.2019.01020
[11] K. Gavrilyuk, R. Sanford, M. Javan and C. G. M. Snoek, "Actor-Transformers for Group Activity Recognition," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 836-845, doi: 10.1109/CVPR42600.2020.00092.
[12] S. Li et al., "GroupFormer: Group Activity Recognition with Clustered Spatial-Temporal Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 13648-13657, doi: 10.1109/ICCV48922.2021.01341
[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90
[14] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997; 9 (8): 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735
[15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning - Volume 37 (ICML'15), Vol. 37. JMLR.org, 448–456.
[16] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.