On the Development of a Classification Based Automated Motion Imagery Interpretability Prediction

Hua-mei Chen[1], Genshe Chen[1](✉), and Erik Blasch[2]

[1] Intelligent Fusion Technology, Inc., Germantown, MD 20876, USA
gchen@intfusiontech.com
[2] MOVEJ Analytics, Dayton, OH, USA

Abstract. Motion imagery interpretability is commonly represented by the Video National Imagery Interpretability Rating Scale (VNIIRS), which is a subjective metric based on human analysts' visual assessment. Therefore, VNIIRS is a very time-consuming task. This paper presents the development of a fully *automated motion imagery interpretability prediction*, called AMIIP. AMIIP employs a three-dimensional convolutional neural network (3D-CNN) that accepts as inputs many video blocks (small image sequences) extracted from motion imagery, and outputs the label classification for each video block. The result is a histogram of the labels/categories that is then used to estimate the interpretability of the motion imagery. For each training video clip, it is labeled based on its subjectively rated VNIIRS level; thus, the required human annotation of imagery for training data is minimized. By using a collection of 76 high definition aerial video clips, three preliminary experimental results indicate that the estimation error is within 0.5 VNIIRS rating scale.

Keywords: Motion imagery interpretability · VNIIRS · National Imagery Interpretability Rating Scale · Deep learning · 3D-CNN

1 Introduction

The pervasive use of still and video imagery from advanced imaging sensors and computing systems produces a desire to quantify the interpretability of imagery. Hence, the Video National Imagery Interpretability Rating Scale (VNIIRS) has been developed. The VNIIRS defines different levels of interpretability based on the types of tasks an analyst can perform with videos of a given VNIIRS rating. The VNIIRS concept assists imagery analysts to perform demanding interpretation tasks as the quality of the imagery increases. Users of motion imagery exploit the interpretability of motion imagery as a guide to determine its relevance value. However, the availability of increasing volumes of motion imagery data makes it infeasible to rely on human analysts for rating all the motion imagery.

The VNIIRS standard is documented in *Motion Imagery Standards Board* (MISB) Standard 0901.2 [1]. A concurrent recommended practice MISB RP 1203.3 describes

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12668, pp. 75–88, 2021.
https://doi.org/10.1007/978-3-030-68793-9_6

two equations, including 1) a video quality equation that predicts the overall appearance of the video and 2) the VNIIRS interpretability estimation equation, or *Motion Imagery Quality Equation* (MIQE) [2], that predicts the VNIIRS rating of a given video based on resolution, blur, noise, camera/platform motion, overall contrast, foreground contrast and motion, and artifacts. The difference between task-based interpretability and appearance-based video quality is defined by ITU-T Recommendation P. 912 [3]. MIQE has been previously used for the development of an automated VNIIRS assessment system [4, 5] as well as the General Image Quality Equation (GIQE) [6, 7]; however, experience has shown that measurements of interpretability from engineering metrics are not easily reconciled with measurements of interpretability from human analysts [8, 9].

The *value of aerial motion imagery* depends on the resolution, quality, and intended use. Standard aerial imagery typically involves a single camera pointing towards a region of interest from which to determine the dynamic content. The interpretability of such aerial imagery compounds for situations such as wide-area motion imagery (WAMI) where multiple cameras are collocated and the images are stitched together to increase the field of view. WAMI exploitation from aerial collects includes multi-object detection [10], coordinated target association [11], distributed processing [12], multi-object tracking [13], and image mosaicking [14]. The ability for aerial imagery to support operational needs of registration, detection, recognition, classification, and identification requires intelligent methods for processing. Examples inherent in any imagery processing pipeline is image compression with effects of image quality [15], interpretability degradation [16], and multiview reconstruction [17]. Additional motion imagery developments of semantic labeling [18] and tensor methods [19] enable fusion on different types of aerial imagery. For example, synthetic aperture radar (SAR) for moving targets [20] utilizes the NIIRS, but can be enhanced for a VNIIRS and advancements in deep learning [21].

This paper describes a fully automated VNIIRS estimation approach without resorting to MIQE. The idea is to cast VNIIRS estimation as a video classification problem and develop an advanced machine learning (ML) 3D convolutional neural networks (CNNs). To realize the ML approach, the video clip is segmented into many short, small video blocks (VBs). The classification result is a histogram of predicted labels/categories that can be used to estimate the VNIIRS level of the test video clip.

This paper is organized as follows. In Sect. 2, the VNIIRS standard is reviewed. Section 3 present works that are related to the present study. The *automated motion imagery interpretability prediction* (AMIIP) approach is detailed in Sect. 4, followed by experimental results in Sect. 5. Conclusions are provided in Sect. 6.

2 Video National Imagery Interpretability Rating Scale and Motion Imagery Quality Equation

Measures of visual interpretability are used in various ways: [1]

- By users to describe a user's visual interpretability needs
- By mission planners in predictive equations
- By users to measure visual interpretability of collected images
- By developers to assess sensor design and image interpretability

The VNIIRS is designed to quantify the interpretability of motion imagery through a set of pre-defined criteria for seven orders of battle or content domains[1]. Each of the written criteria contains the following five components: Analyst Task (such as 'track' or 'confirm'), Object of Interest, Associated Activity or Behavior, Environment, and optional Object Reference Examples. We refer to the criteria as the *VNIIRS components* in this paper. For instance, the following criterion is provided to define the level 5 VNIIRS rating in CULTURE content domain: "*Track the movement of - a car, SUV, van, or light truck- driving independently - on roadways in medium traffic - (mid & full-size cars & trucks:* 5 m – 6 m *length).*" The VNIIRS has interpretability levels ranging from 3 to 11, and each content domain has at least one task defined for each of the nine VNIIRS level.

Both sensor design parameters and imaging conditions affect the VNIIRS rating of Motion Imagery. For example, the sensor design parameters are relative edge response (RER), signal to noise ratio (SNR), *peak SNR* (PSNR) resolution or frame size, and compression are sensor design parameters; while the imaging conditions are *ground sample distance* (GSD), atmospheric conditions, target illumination, object and camera motion, and amount of clutters. Attempts have been made to objectively predict the VNIIRS based on these factors, and the result is *motion imagery quality equation*, or MIQE [2], which expresses the instantaneous interpretability estimate for the k_{th} frame as a function of resolution (in term of ground sampling distance, GSD), blur (in terms of relative edge response, RER), noise (in terms of peak signal to noise ratio in dB, PSNR), camera/platform motion, overall contrast, foreground contrast and motion, and artifacts as:

$$I_k = 14 - \log_2 GSD_k - \log_2\left({}^{1}\!/_{RER_k}\right) - \exp(0.5 \cdot (PSNR_c - PSNR_k)) \qquad (1)$$

$$-\Delta I_{camara} - \Delta I_{contrast} - \Delta I_{movers} - \Delta I_{artifacts}$$

where PSNR_c is the critical point which has been experimentally determined to be 26 dB. Interesting readers are referred to [3] for the detailed definition of each involved variable as well as the recommended implementations.

3 Related Work

The current study is related to deep learning-based no-reference video quality assessment (NR-VQA). The goal of NR-VQA is to estimate the mean opinion score (MOS) of the video. In [3], video quality is formally defined as a metric of five levels: 1. Bad; 2. Poor; 3. Fair; 4. Good; and 5. Excellence. Nevertheless, other standardized quality ratings also exist, such as a continuous scale ranging from 1.0 to 100.0, but Huynh-Thu *et al.* [22] noted that there are no statistical differences between the different scales used for the same visual stimuli.

Among varies NR-VQA schemes [23–27], the work proposed by Varga [28] mostly resembles the work presented in this paper in the sense that both works cast video quality or interpretability prediction as a classification problem. Nevertheless, there are

[1] This is based on the standard MISB ST 0901.2. However, in the newest standard MISB ST 0901.3, criteria are defined for three orders of battle.

several differences. 1) In [28], Two-dimensional deep convolutional neural networks are employed to extract frame-wise feature vectors followed by varies element-wise fusion strategies to form video-level feature vector, while AMIIP utilizes a 3D CNN for feature extraction. 2) In [28], there is only one classification task involved in each video clip; while in the AMIIP approach, a large number of small video blocks (VBs) are extracted from each video clip and classification is performed on each VB. Finally, 3) in [28] a support vector regressor is employed to map the temporally pooled video-level feature vectors to perceptual quality scores, while AMIIP employs a simple linear classifier for classifying the encoded feature vector for each VB.

Another closely related area is video analysis through spatiotemporal features such as action recognition and content-based video classification. The spatiotemporal features in [29] for action recognition are learned by performing 3D convolutions on an image block of size 60 (H) $\times$ 40 (W) $\times$ 33 (D) which in turn is obtained by stacking five channels[2] of the input video block of size 60 (H) $\times$ 40 (W) $\times$ 7 (F). In [30] for content-based video classification, different strategies for extending the connectivity of a CNN in time domain are proposed and compared, including Late Fusion, Early Fusion and Slow Fusion, in addition to the base line Single Frame. In [31], a general-purpose spatiotemporal feature learning scheme, known as C3D, is proposed. C3D features the use of small $3\times3\times3$ convolution kernels in all layers making it an appealing scheme for end-to-end deep learning applications. The AMIIP deep convolutional neural network is based on the C3D network structure with some modifications to accommodate the specific input sizes for the application. Specifically, AMIIP doubles the number of convolutional layers from Conv3a to Conv5b and adds Conv6 in the network architecture. Due to the deeper CNN architecture, to avoid the vanishing gradient problem [32], the concept of residual blocks is employed to build Conv3a to Conv5b network blocks. The details of the 3D CNN architecture are illustrated in Fig. 2. The imagery analysis is to maintain situation awareness [33] through an interpretability index demonstrating classification performance.

4 Methodology

The workflow of the proposed automated motion imagery interpretability prediction (AMIIP) method is shown in Fig. 1 which consists of three stages: Video Block (VB) Generation, 3D-CNN Based VB Classification, and VNIIRS Prediction. In the first stage of *VB Generation*, a large number of small video blocks are generated from the center part of the video, which is the part of the video that analysts focus on. In the second stage of *VB Classification*, each video block is fed into a 3D-CNN based classifier for label prediction. The result is a frequency histogram, which is then used for VNIIRS prediction at the third stage for *VNIIRS Prediction*. The details of each stage are provided next.

4.1 Video Block Generation

As shown in Fig. 1, the first stage of the proposed approach is video block generation. The output of this stage is a large number of fixed sized image volumes employed as

[2] The five channels are defined as gray, gradient-x, gradient-y, optflow-x and optflow-y.

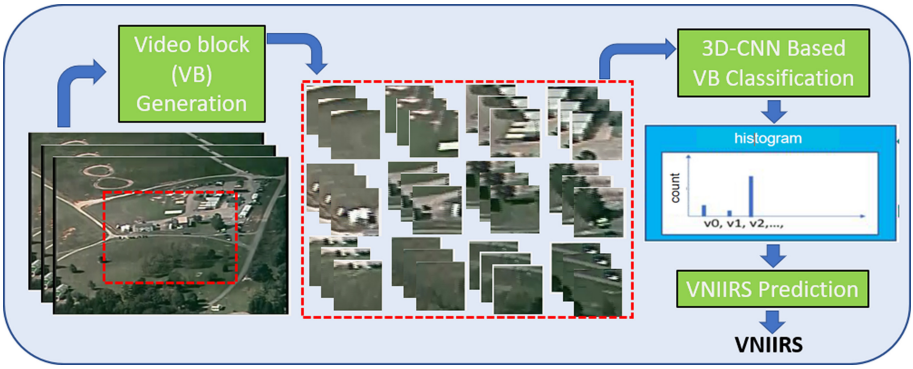


Fig. 1. Overview of the proposed automated motion imagery interpretability prediction (AMIIP) approach with (1) video block generation, (2) video block classification, and (3) VNIIRS prediction based on aggregate video block processing

inputs to the subsequent 3D CNN. There are at least two reasons that support the use of small video blocks instead of using the entire video. 1) Motion is an important cue for an image analyst to subjectively assign a VNIIRS level. To capture the information pertaining to motion, a reasonable number of consecutive frames need to be grouped instead of just three frames as employed by Tran *et al*. in [31] or seven frames by Ji *et al*. in [29]. As a result, the spatial extent has to be small enough in order to have a sufficient number of input data in one batch. 2) Like visual quality analysis of video, interpretability of motion imagery should be content independent. By keeping the spatial extent of video blocks from being too large, it is less likely that an object completely resides in a video block. In this work, different video block sizes are experimented. Specifically, the values tested of {height(h) $\times$ width(w) $\times$ frames(f)} include {64 $\times$ 64 $\times$ 16}, {32 $\times$ 32 $\times$ 16}, and {64 $\times$ 64 $\times$ 32}. In our implementation, AMIIP applies a 3D sliding window approach at the center part of each frame with step sizes Δh, Δw, Δf equal to h, w, and $0.5f$ respectively.

Intuitively, an ideal VB should contain sufficient spatial (in x-y plane, or frame-wise) and temporal (in the z direction) variations. A frame without spatial variation is an indication of lack of foreground in the frame, and a volume without intensity variation along the z-direction which indicates either no object is present, or the object is not moving. For this reason, two VB selection criteria are devised to select VBs of large spatial and temporal variations. In the first criterion, *spatial STD test*, AMIIP first computes the standard variation of pixel intensities for each frame of a VB. Next, AMIIP denotes the median of the STDs of all frames in a VB as $\delta_{spatial}$. Then, a selected VB must satisfy:

$$\delta_{spatial} > Th_{spatial} \tag{2}$$

Likewise, in the second criterion, *temporal STD test*, AMIIP first computes the standard deviation of pixel intensities for each x-y position along z-direction for each VB. Next, AMIIP denotes the 99th percentile of the STDs of all x-y positions in a VB as $\delta_{temporal}$. Then, a selected small volume must satisfy:

$$\delta_{temporal} > Th_{temporal} \tag{3}$$

To test the validity of the two VB selection criteria, experiments are conducted to compare the performances obtained by selecting the VBs that satisfied both criteria and by selecting the VBs that do not satisfy both criteria. In other words, in the latter case, the selected VBs are either homogeneous in each frame or almost no moving objects are contained in them.

4.2 Three-Dimensional CNN Based Video Block Classification

Inspired by the C3D network structure shown in Fig. 2(a), AMIIP is designed by constructing several variants of the C3D based on the size of the input video block. One typical structure is given in Fig. 2(b), which corresponds to input block size of 64 × 64 × 32. In this case, because of the longer input image sequence, conv6 and the subsequent max pooling layers are added to reduce the dimension of the feature vector. In addition, batch normalization, dropout and residual blocks are incorporated into the proposed AMIIP 3D CNN structure. The final feature vector has length 256 and the number of class is 9, which is explained in Sect. 5. For other variants of C3D, for example, when the input video block size is 64 × 64 × 16[3], conv6 related layers such as bn6, relu6, and maxPool6 are removed.

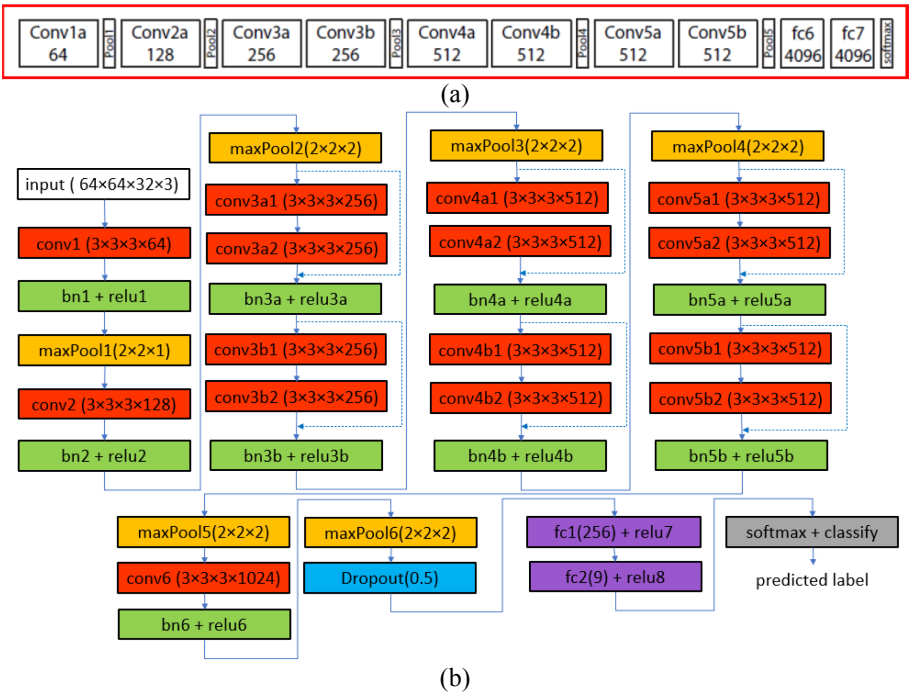


Fig. 2. (a) Basic C3D structure for video of size 112 × 112 × 16. (b) The proposed 3D CNN structure for video block of size 64 × 64 × 32.

[3] Different video block sizes are experimented in this paper.

4.3 Video NIIRS Prediction from Label Histogram

Results of the second stage for each test video clip can be represented as a label histogram as illustrated in Fig. 3, where each label in the horizontal axis corresponds to a different VNIIRS level. To obtain the predicted VNIIRS level for the test video clip, AMIIP simply computes the weighted average of the nearest 5 entries surrounding the highest histogram.

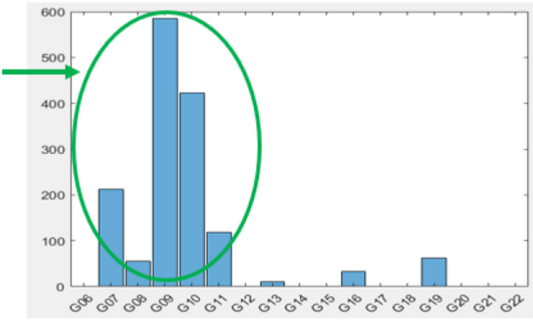


Fig. 3. VNIIRS prediction from label histogram

5 Experiments

As a preliminary study, in this section we report three experimental results using a set of 76 high definition aerial video clips among which 66 clips are used as the training data set and the remaining 10 video clips are adopted as test data set. All clips have been subjectively assigned non-integer VNIIRS levels ranging from 7 to 11 by several image analysts. Some information about the video clips including the duration, frame-size and frame-rate are provided in Table 1. The AMIIP classifiers are implemented in MATLAB. In all experiments, the batch size is 54 with stochastic gradient descent optimization with initial learning rate 0.001 and momentum 0.8.

Table 1. Information of video clips used in the experiments.

Clip length	Frame size (width × height)	Frame rate (fps)
10 s	1920 × 1080	25

5.1 Data Preparation

We first assign each training video clip a group label from G14 to G22 according to its ground truth VNIIRS levels rounded to half-integers. For example, the group G16 includes the training clips whose rounded VNIIRS values are 8. In other words, the ground truth VNIIRS values of the clips in the group G16 are between 7.75 and 8.25. In this way, we divide the 66 training clips into nine groups of different labels.

5.2 Experiment 1: Performance Comparison of Two Spatial Extents

In the experiment to compare *spatial extent*, we evaluate the performances of two different VB sizes, 64 × 64 × 16 and 32 × 32 × 16. AMIIP first generates VBs of size 64 × 64 × 16 and select 22984 VBs by using both the spatial and the temporal STD tests given in Eqs. (2) and (3). Both thresholds are set to be 10. For VBs of size 32 × 32 × 16, in order to use the same 3D CNN structure, AMIIP re-uses the generated VBs of size 64 × 64 × 16 by dividing each VB into four 32 × 32 × 16 VBs followed by up-sampling each 32 × 32 × 16 VB to 64 × 64 × 16 VB. The sampling procedure, as illustrated in Fig. 4, also ensures that exactly the same training VBs are involved in both cases. The same procedure is employed when preparing the VBs of test clips. Figure 5 shows the results of the 10 test video clips. The numerical results are provided in Table 2. Clearly, VBs of size 64 × 64 × 16 outperform VBs of size 32 × 32 × 16 in this experiment.

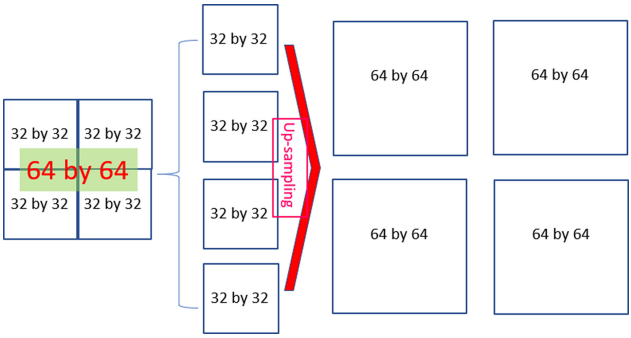


Fig. 4. Video Blocks generation in experiment 1 (spatial extent)

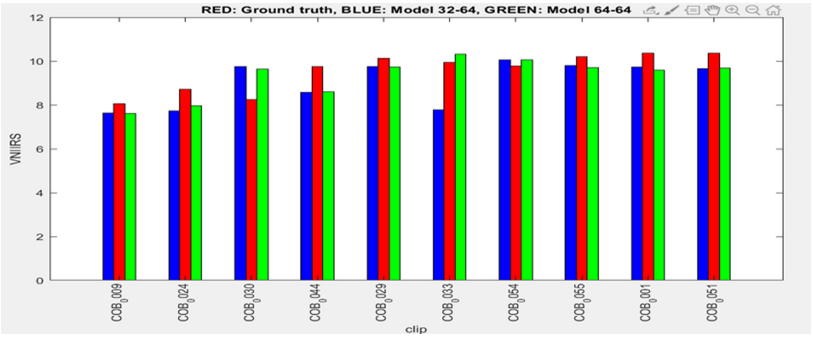


Fig. 5. Result of experiment 1. Blue: VBs of size 32 × 32× 16; Red: Ground truth; Green: VBs of size 64 × 64× 16. (Color figure online)

Table 2. Numerical result of experiment 1

Video Block	Total # of VBs	Mean error	STD
64 × 64 × 16	22984	0.67	0.35
32 × 32 × 16	91936 (22984 × 4)	0.86	0.34

5.3 Experiment 2: Performance Comparison of Two Temporal Extents

In the experiment to compare *temporal extent*, we evaluate the performances of two different VB lengths, 64 × 64 × 16 and 64 × 64 × 32. AMIIP first generates VBs of size 64 × 64 × 32 and select 23308 VBs by using both the spatial and the temporal STD tests given in Eqs. (2) and (3). Both thresholds are set to be 10. For VBs of size 64 × 64 × 16, in order to use exactly the same training data, AMIIP re-uses the VBs of size 64 × 64 × 32 by splitting each VB into two 64 × 64 × 16 VBs. This procedure is graphically illustrated in Fig. 6. Due to the different VB lengths, different 3D CNN structures are used as explained in Sect. 4.2. The experiment is repeated three times, resulting in three classifiers for each VB size. In addition, during the test phase, instead of using all VBs, AMIIP randomly selects 40 and 100 VBs per clip in order to speed up the prediction process. The same test VBs are used in all three runs. The results are provided in Fig. 7 and Table 3. From the results, we observe: 1) VBs of size 64 × 64 × 32 outperform VBs of size 64 × 64 × 16; 2) Selecting 100 VBs for each test clip outperforms selecting 40 VBs for each test clip; and 3) even when the same training VBs are used, performances of the three trained classifiers fluctuates. The first observation seems indicate that longer VBs captures motion information better than shorter VBs, while the second observation suggests that an insufficient number of test VBs per video clip deteriorates the performance.

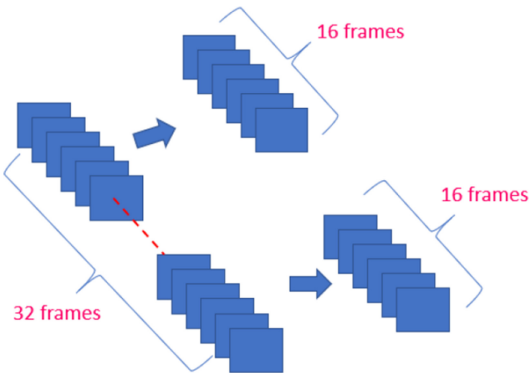
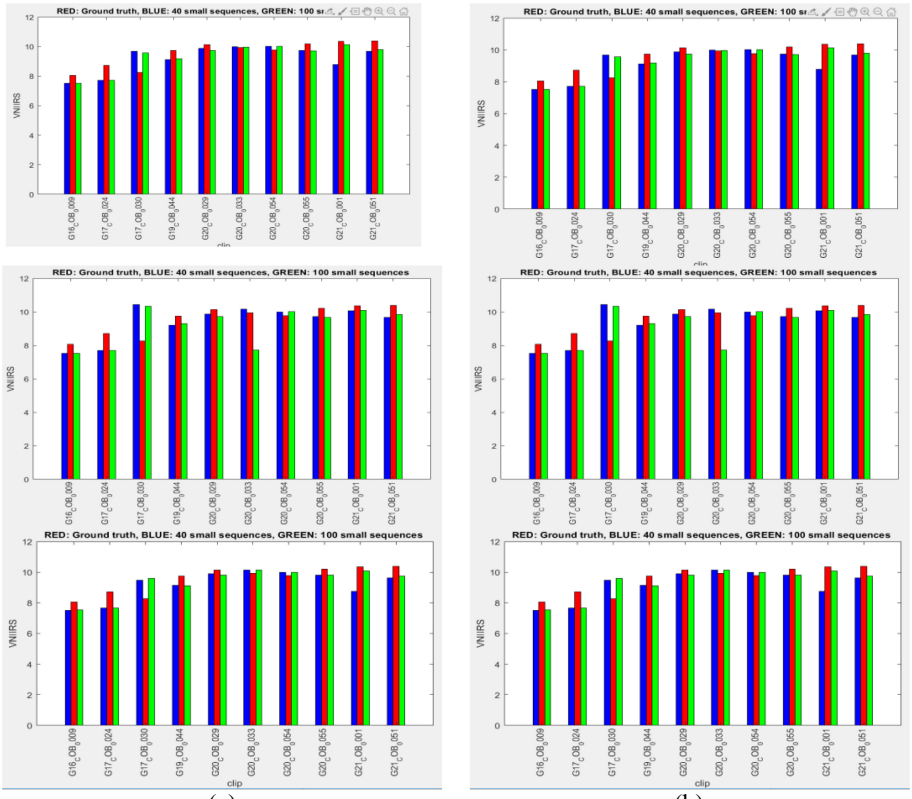


Fig. 6. Video Block generation scheme adopted in experiment 2 (temporal extent)

Table 3. Numerical results of three runs of the second experiment *(temporal extent)*.

VB length	# of VBs	Avg. error (40 VBs)	Avg. STD (40 VBs)	Avg. error (100 VBs)	Avg. STD (100 VBs)
32-1	23308	0.650	0.694	0.509	0.385
32-2	23308	0.560	0.561	0.410	0.260
32-3	23308	0.709	0.720	0.481	0.371
16-1	46616	0.689	0.503	0.538	0.380
16-2	46616	0.649	0.596	0.830	0.731
16-3	46616	0.681	0.477	0.560	0.378



(a) (b)

Fig. 7. Results of the second experiment (temporal extent). Each row is the results of one run of the experiment. (a) VBs of size 64 × 64 × 16. (b) VBs of size 64 × 64 × 32. Blue: 40 VBs per test clip. Red: Ground truth. Green: 100 VBs per test clip. (Color figure online)

5.4 Experiment 3: Test the Effectiveness of Both VB Selection Criteria

The final experiment tests the effectiveness of the spatial STD test and the temporal STD test for VB selection. Both tests are given in Eqs. (2) and (3). For this experiment, we adopt the setup of Experiment 2 for the case of VBs of size $64 \times 64 \times 32$, but the selected training VBs are those that fail both tests. In addition, we also test the performance of employing more VBs in the test phase. Table 4 and Fig. 8 show the results. In Table 4, the values in the first three rows are taken from the first three rows of Table 3. Surprisingly, although the performance resulting from those training VBs that pass both tests is better, the difference between them is not significant. It indicates that the *VBs that mostly contain homogeneous background and lack of moving objects do capture useful information for VNIIRS prediction*. However, the performances of the resulting classifiers seem not as stable as those resulting from using the classifiers trained by the VBs that pass both criteria.

Table 4. Numerical results of three runs of the third experiment (block selection).

VB selection tests	# of VBs	Avg. error (230 VBs)	Avg. STD (230 VBs)	Avg. error (100 VBs)	Avg. STD (100 VBs)
Pass – 1	23308	n/a	n/a	0.509	0.385
Pass – 2	23308	n/a	n/a	0.410	0.260
Pass – 3	23308	n/a	n/a	0.481	0.371
Fail – 1	18904	0.420	0.340	0.461	0.307
Fail – 2	18904	0.617	0.691	0.897	0.814
Fail – 3	18904	0.423	0.518	0.593	0.561

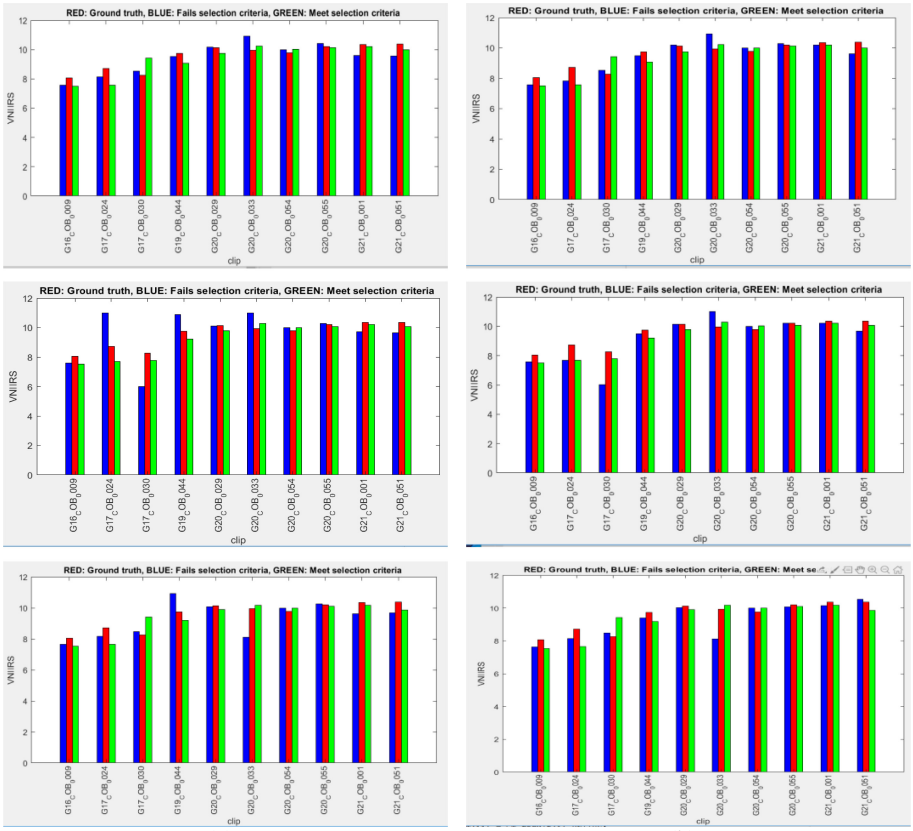


Fig. 8. Results of the third experiment (block selection). Each row is the results of one run of the experiment. (a) 100 test VBs per test clip. (b) 230 test VBs per test clip. Blue: VBs fail to pass both VB selection tests. Red: Ground truth. Green: VBs pass both VB selection tests. (Color figure online)

6 Concluding Remarks

In this paper, a fully automated approach for predicting the interpretability of motion imagery, based on advanced 3D convolutional neural networks is presented. The AMIIP (*automated motion imagery interpretability prediction*) predicts the interpretability of high definition aerial videos with VNIIRS ranging from 7 to 11, by casting it as a video classification problem. Due to the large frame size and the adoption of 3D CNN, AMIIP divides the entire video clip into many small video blocks (VBs) and predicts the VNIIRS level of a test clip based on the labels predicted for all VBs. The AMIIP 3D CNN structure is based on the C3D network that utilizes small 3D convolutional kernels. Using a set of 76 short HD aerial video clips, three preliminary experimental results demonstrate the feasibility of the proposed fully automated VNIIRS prediction.

One surprising observation is that VBs with mostly homogeneous backgrounds still contain information that can be used by the 3D classifier to distinguish clips with different interpretability. However, due to the limited video dataset tested, future work will investigate (1) more video data sets of different lengths, (2) different imagery types, and (3) multimodal analysis. The extent of these variations for robustness is required to consolidate and verify the findings reported in this paper.

References

1. MISB ST 0901.2: Video-National Interpretability Rating Scale, Feb 2014
2. MISB RP 1203.3: Video Interpretability and Quality Measurement and Prediction, Feb 2014
3. ITU-T Recommendation P.912, Subjective Video Quality Assessment Methods for Recognition Tasks, Aug 2008
4. Blasch, E., Kahler, B.: Application of VNIIRS for target tracking. In: Proceedings of SPIE vol. 9473 (2015)
5. Blasch, E., Kahler, B.: V-NIIRS fusion modeling for EO/IR systems. In: IEEE National Aerospace and Electronics Conference (2015)
6. Blasch, E., Chen, H-M., Wang, Z., Jia, B., et al.: Target broker compression for multi-level fusion. In: IEEE National Aerospace and Electronics Conference (2016)
7. Blasch, E., Chen, H-M., Wang, Z., Jia, B., et al.: Compression induced image quality degradation in terms of NIIRS. In: IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (2016)
8. Zheng, Y., Dong, W., et al.: Qualitative and quantitative comparisons of multispectral night vision colorization techniques. Opt. Eng. **51**(8), 08004 (2012)
9. Zheng, Y., Blasch, E., Liu, Z.: Multispectral Image Fusion and Colorization. SPIE Press (2018)
10. Palaniappan, K., et al.: Moving object detection for vehicle tracking in wide area motion imagery using 4D filtering. In: International Conference on Pattern Recognition (ICPR) (2016)
11. Snidaro, L., García, J., Llinas, J., Blasch, E. (eds.): Context-Enhanced Information Fusion. ACVPR, Springer, Cham (2016). https://doi.org/10.1007/978-3-319-28971-7
12. Wu, R., Liu, B., Chen, Y., et al.: A Container-based elastic cloud architecture for pseudo real-time exploitation of wide area motion imagery (WAMI) stream. J. Signal Process. Syst. **88**(2), 219–231 (2017)
13. Al-Shakarji, N.M., Bunyak, F., Seetharaman, G., Palaniappan, K.: Robust multi-object tracking for wide area motion imagery. In: IEEE Applied Imagery Pattern Recognition Workshop (AIPR) (2018)
14. Aktar, R., AliAkbarpour, H., Bunyak, F., Seetharaman, G., Palaniappan, K.: Performance evaluation of feature descriptors for aerial imagery mosaicking. In: IEEE Applied Imagery Pattern Recognition (AIPR) Workshop (2018)
15. Zheng, Y., Chen, G., Wang, Z., et al.: Image quality (IQ) guided multispectral image compression. In: Proceedings of SPIE, vol. 9871 (2016)
16. Blasch, E., et al.: Prediction of compression-induced image interpretability degradation. Opt. Eng. **57**(4), 043108 (2018)
17. Gao, K., Yao, S., AliAkbarpour, H., Agarwal, S., Seetharaman, G., Palaniappan, K.: Sensitivity of multiview 3D point cloud reconstruction to compression quality and image feature detectability. In: IEEE Applied Imagery Pattern Recognition (AIPR) Workshop (2019)
18. Al-Shakarji, N.M., Bunyak, F., AliAkbarpour, H., Seetharaman, G., Palaniappan, K.: Performance evaluation of semantic video compression using multi-cue object detection. In: IEEE Applied Imagery Pattern Recognition (AIPR) Workshop (2019)

19. Prasath, V.B.S., Pelapur, R., Seetharaman, G., Palaniappan, K.: Multiscale structure tensor for improved feature extraction and image regularization. IEEE Trans. Image Process. **28**(12), 6198–6210 (2019)
20. Çetin, M., Stojanović, I., Önhon, N.O., Varshney, K., Samadi, S., et al.: Sparsity-driven synthetic aperture radar imaging: reconstruction, autofocusing, moving targets, and compressed sensing. IEEE Signal Process. Mag. **31**(4), 27–40 (2014)
21. Majumder, U., Blasch, E., Garren, D.: Deep Learning for Radar and Communications Automatic Target Recognition. Artech House, Norwood (2020)
22. Huynh-Thu, Q., Garcia, M.N., Speranza, F., et al.: Study of rating scales for subjective quality assessment of high-definition video. IEEE Trans. Broadcast **57**(1), 1–14 (2011)
23. Zhang, Y., Gao, X., He, L., et al.: Blind video quality assessment with weakly supervised learning and resampling strategy. IEEE Trans. Circuits Sys. Video Tech. **29**(8), 2244–2255 (2018)
24. Li, Y., et al.: No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Trans Circuits Sys Video Tech. **26**(6), 1044–1057 (2016)
25. Shahid, M., Rossholm, A., Lövström, B., Zepernick, H.-J.: No-reference image and video quality assessment: a classification and review of recent approaches. EURASIP J. Image Video Process. **2014**(1), 1–32 (2014). https://doi.org/10.1186/1687-5281-2014-40
26. Vega, M.T., Sguazzo, V., Mocanu, D.C., et al.: An experimental survey of no-reference video quality assessment methods. Int. J Pervasive Comp. Comm **12**(1), 66–86 (2016)
27. Xu, L., Lin, W., Kuo, C.C.J.: Visual quality assessment by machine learning. Springer, Berlin (2015)
28. Varga, D.: No-reference video quality assessment based on the temporal pooling of deep features. Neural Process. Lett. **50**(3), 2595–2608 (2019)
29. Ji, S., Xu, W., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
30. Karpathy, A., Toderici, G., et al.: Large-scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
31. Tran, D., Bourdev, L., et al.: Learning spatiotemporal features with 3D convolutional networks. IEEE International Conference on Computer Vision (2015)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
33. Blasch, E., Seetharaman, G., et al.: Wide-area motion imagery (WAMI) exploitation tools for enhanced situation awareness. In: IEEE Applied Imagery Pattern Recognition Workshop (2012)