Statistics of Optical Flow

A Project Report

submitted by

Pavan C M (13990)

as part of

Digital Video: Perception and Algorithms (E9 206)

undertaken during

August - December 2017

DEPARTMENT OF ELECTRICAL COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF SCIENCE BANGALORE - 560012

December 11, 2017

TABLE OF CONTENTS

CHAPTER 1
Introduction

1.1 Optical Flow

Optical Flow represents apparent motion of edges, surfaces and objects caused mainly due to relative motion between the observer and scene [1]. In a dynamic scene this can occur at any or every location of the scene. It mainly refers to displacement associated with intensity patterns. Optical flow has major applications in motion analysis.

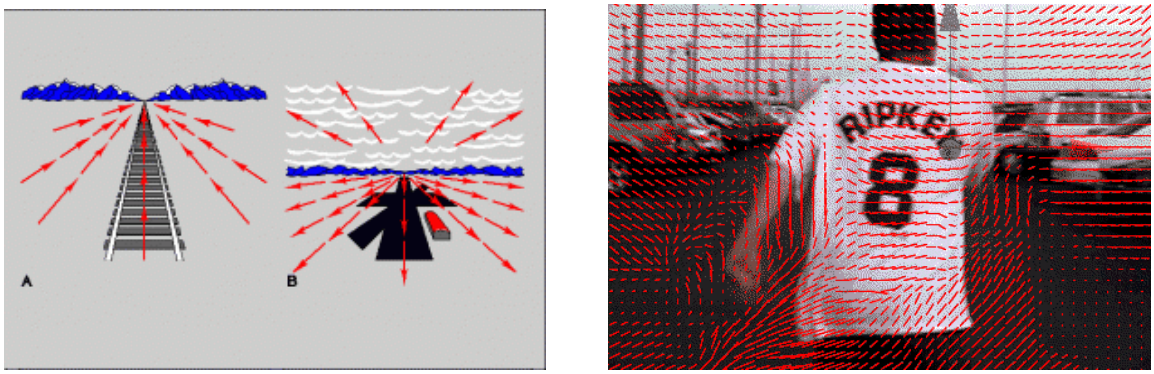


Figure 1.1: Figures illustrating optical flow using needle diagram [2],[3]

Finding optical flow has been a major area of research for nearly three decades. One of the first methods to address optical flow calculations were proposed by Horn and Schunk [4]. Lucas and Kanade [5] proposed another simpler method which assumed constant optical flow in a neighborhood and was based on least squares solution. Further improvements to Horn-Schunk method was proposed by Black and Anandan [6] by employing robust statistics. Brox *et al.* [7] proposed non-linear optical flow along with spatio-temporal smoothing constraint to obtain more accurate and smoother estimates. Parallel to the above methods, Fleet and Jepson [8] came up with perceptual optical flow model which used phase information for optical flow calculation. Recently due to advent deep learning methods, certain learning algorithms like Flownet [9] and Flownet-2 [10] based on Convolutional Neural Network have been successfully demonstrated to perform better when compared with classical methods. In this project all analysis is done by considering Classic-NL [11] method for optical flow calculation due to it's simplicity, speed and better accuracy (it's accuracy might not be as good as that of learning based methods). The results reported here may not be easily reproducible for all optical flow algorithms as the results are very much tied to the optical flow estimates.

1.2 Blind Video Quality Assessment

Video Quality Assessment (VQA) is an emerging area of research which aims at establishing objective quality evaluation methodologies for predicting video quality. Objective quality assessment algorithms have several applications like real-time monitoring of quality in networks, improve quality of experience (QoE) for users, evaluating compression algorithms etc. Objective quality assessment methods fall into three categories: 1) full-reference (FR), 2) reduced-reference (RR), and 3) blind or no-reference (NR) approaches. In this project only NR approach is considered where reference video or pristine video is not available during evaluation phase. The methods discussed in this report is mainly based on Video BLIINDS algorithm [12]. A brief description of features used in video BLIINDS is provided here.

1.2.1 Motion Coherency

Motion silencing phenomenon due to flicker is characterized by a structure tensor

$$\begin{aligned} S &= \begin{pmatrix} f(M_x) & f(M_x, M_y) \\ f(M_x, M_y) & f(M_y) \end{pmatrix} \\ \text{where} \quad f(V) &= \sum_{l,k} w[i,j]V(i-l, j-k) \end{aligned} \tag{1.1}$$

and $M_x(i,j)$ and $M_y(i,j)$ are horizontal and vertical motion vectors at pixel (i,j) respectively and w is a window around the neighborhood of (i,j). Relative discrepancy of eigen values of S is a measure of motion biased in a particular direction. For eigen values λ_1 and λ_2 motion coherence measure is given by

$$C = \left(\frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}\right)^2 \tag{1.2}$$

1.2.2 Camera Motion

Camera motion is characterized by analyzing the variation of motion vectors.

$$\begin{aligned} M &= mode_{i=1,\ldots,m}(\sqrt{(M_{X(i)})^2 + (M_{Y(i)})^2}) \\ E &= \frac{1}{m}\sum_{i=1}^{m}\left(\sqrt{(M_{X(i)})^2 + (M_{Y(i)})^2}\right) \end{aligned} \tag{1.3}$$

where m is the number of motion vectors per frame. The quantities M and $|E-M|$ are then averaged over the frames of video resulting in M_{ave} and $|E-M|_{ave}$. Global motion characterization measure is given by

$$G = \frac{|E-M|_{ave}}{1 + M_{ave}} \tag{1.4}$$

1.2.3 Spatio-Temporal DCT Model

Frame differences is computed for every consecutive frame and each difference frame is partitioned into $n \times n$ patches(BLIINDS algorithm employs $n = 5$ value). The 2D DCT is then applied to every $n \times n$ patch. The histogram of each frequency coefficient from all $n \times n$ patches in each difference frame is fitted with a Generalized Gaussian Model (GGD) described in equations 1.5, 1.6, 1.7.

$$f(x|\alpha, \beta, \gamma) = \alpha e^{-(\beta(x-\mu))^{\gamma}} \tag{1.5}$$

where μ is the mean, γ is the shape parameter, and α and β are normalizing and scale parameters given by

$$\alpha = \frac{\beta\gamma}{2\Gamma(1/\gamma)} \tag{1.6}$$

$$\beta = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}} \tag{1.7}$$

where σ is the standard deviation, and Γ denotes the ordinary gamma function

$$\Gamma(z) = \int_0^{\infty} t^{z-1}e^{-t}\,dt \tag{1.8}$$

In order to capture the spectral signatures of videos (pristine and distorted), each $n \times n$ matrix of shape-parameters per difference frame is partitioned into three sub-bands and geometric mean of shape parameter is computed across each band to obtain the final features. More detailed description on spatio-temporal feature extraction is provided at [12].

1.2.4 Spatial Statistics

For capturing spatial distortions Natural Image Quality Indicator (NIQE) [14] is employed for every frame of the video. The features extracted from the above described methods are concatenated and used for training a linear kernel Support Vector Regressor (SVR). Implementation described in [16] is used to conduct quality score prediction.

1.3 Video Quality Database

Analysis and evaluation of algorithms discussed in this project are done with regard to LIVE VQA Database [17]. The LIVE VQA database has a total of 160 videos derived from 10 pristine videos having a rich variety of spatial and temporal content. The database contains videos distorted by four types of distortions: 1) MPEG-2 compression, 2) H.264 compression, 3) wireless distortions, and 4) IP distortions. For evaluation purposes the 80% of the dataset was used for training and the rest for testing with no overlapping content.

CHAPTER 2
Observations and Results

2.1 Spatial Statistics

Optical flow was computed for every video present in the LIVE Video quality database using Classic-NL algorithm. Since videos are have spatial as well as temporal variations, their corresponding optical flow are spatio-temporal in nature as well. The statistics corresponding to spatial and temporal optical flow variations are analyzed separately in this section. The analysis of spatial statistics is motivated from [13].

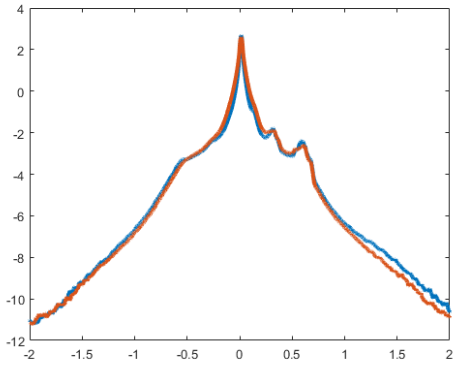
2.1.1 Global Velocity Statistics

Optical flow is characterized by both horizontal and vertical motion vectors computed at every location for pair of frames. Mean subtracted contrast normalized coefficients (MSCN) as given in equation 2.1 are computed for every frame of horizontal and vertical motion field.

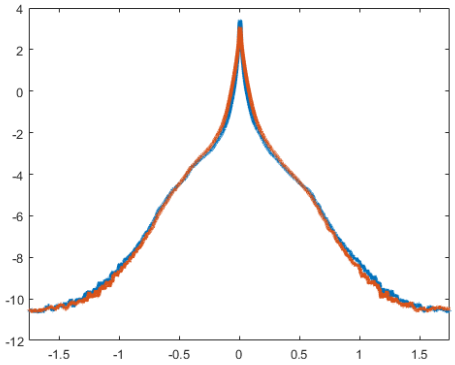
$$\begin{aligned} \hat{I} &= \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \\ \mu(i,j) &= \sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} I_{k,l}(i,j) \\ \sigma(i,j) &= \sqrt{\sum_{k=-K}^{K} \sum_{l=-L}^{L} w_{k,l} (I_{k,l}(i,j) - \mu(i,j))^2} \end{aligned} \tag{2.1}$$

where $i \in 1, 2, ...M, j \in 1, 2, ...N$ are spatial indices, M, N are height and width of motion field respectively, $C = 1$ is constant preventing instabilities when denominator tends to zero, $w = \{w_{k,l} | k = -K, ...K, l = -L, ...L\}$ is a 2D circularly symmetric Gaussian weighting function. MSCN coefficients are calculated for every frame and all coefficients are pooled together to obtain marginal distribution. Figure 2.1 shows log histogram of MSCN coefficients of reference as well as distorted version pooled from all frames of the optical flow. It is evident from the figure there exists no visibly distinguishable property for discriminating the distributions of reference and distorted one. Another observation is that the histograms of optical flow MSCN change with the content of the video which can be observed in figure 2.1. Since the LIVE database contains videos with multiple distortions, the effect of different distortions on MSCN distributions is illustrated in figure 2.2. It can be inferred from the figure that multiple distortions are not distinguished by MSCN coefficients as all distributions tend to overlap although they are perceptually different.

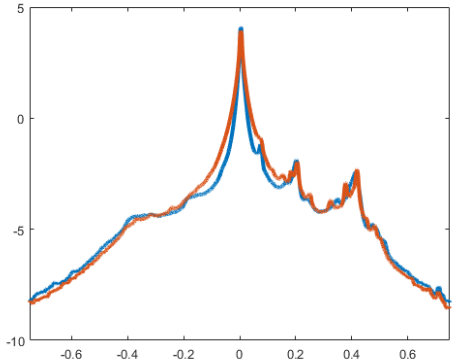
Statistics of a single frame of one of the videos from the database is illustrated in figure 2.3. Although the differences between the pristine and distorted video are easily visible, the same is not being captured by MSCN statistics. Therefore these observations suggest that global statistics of MSCN coefficients aren't particularly effective in capturing the perceived distortions.



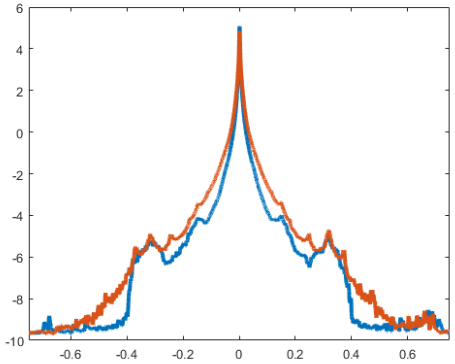
(a) Horizontal velocity (10^{th} video)



(b) Vertical velocity (10^{th} video)

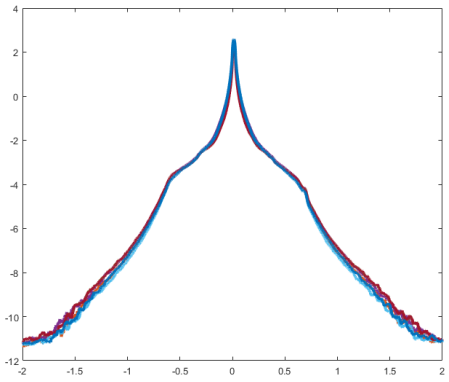


(c) Horizontal velocity (8^{th} video)

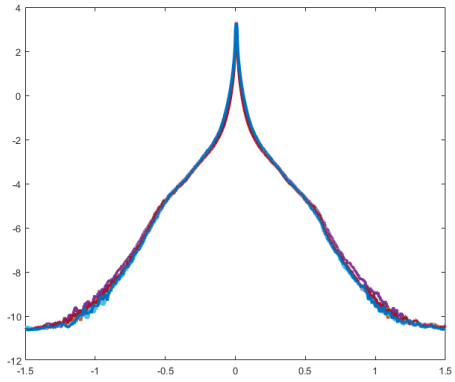


(d) Vertical velocity (8^{th} video)

Figure 2.1: Log Histogram of MSCN coefficients. Blue represents reference video and orange MPEG-2 distortion



(a) Horizontal velocity (3^{rd} video)



(b) Vertical velocity (3^{rd} video)

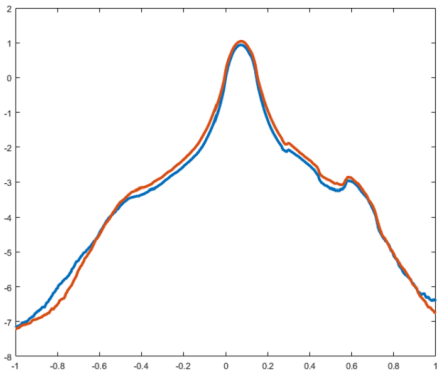
Figure 2.2: Log Histogram of MSCN coefficients depicting all distortions

(a) Pristine - Horizontal Flow Field (10^{th} video)

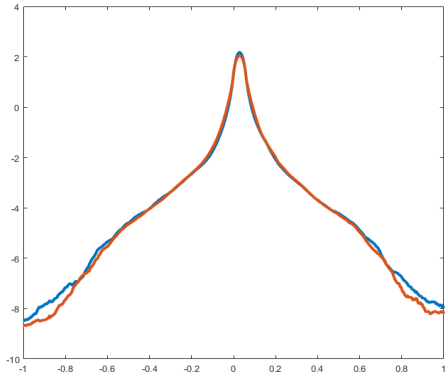
(b) Pristine - Vertical Flow Field (10^{th} video)

(c) Distortion (Wireless) - Horizontal Flow Field (10^{th} video)

(d) Distortion (Wireless) - Vertical Flow Field (10^{th} video)



(e) Horizontal velocity (10^{th} video), blue - reference, orange - distorted



(f) Vertical velocity (10^{th} video), blue - reference, orange - distorted

Figure 2.3: Statistics of a single frame

2.1.2 Local Velocity Statistics

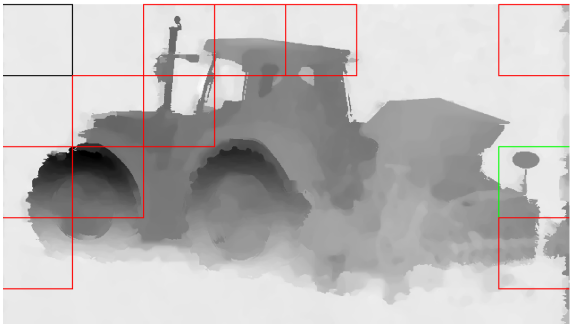
In the previous section it was inferred that global statistics failed in capturing sufficient information about the distortions. Therefore a local approach where every frame is divided into $P \times P$ patches was experimented. Only a subset of the obtained patches is used for quality assessment as not all patches equally contain percieved distortions. Patch selection is done using a similar method explained in NIQE [14]. Although NIQE itself can be employed for quality prediction it was not used as optical flow field frames weren't exactly resembling natural images (appeared to be more synthetic). The variance field obtained in equation 2.1 is a rich source of structural and texture information. Letting the $P \times P$ sized patches be indexed $a = 1, 2, .., A$, a sharpness measure is calculated as

$$\delta(a) = \frac{1}{P^2} \sum_{i,j \in \text{patch a}} \sum \sigma(i,j) \tag{2.2}$$

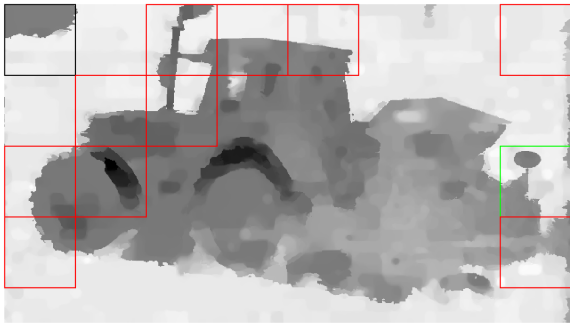
where $\delta(a)$ is simply the local mean of variance field of the patch. Once the sharpness measure is calculated, patches corresponding to top $T\%$ of the sharpness measure are selected to collectively represent the frame. In the above experiments $P = 96$ and $T = 35$ values were used. Figure 2.4 illustrates distribution of patches. It is evident from the figure that local patch coefficients do capture some information with regard to distortion. Similar observations were observed in case of vertical motion field as well. The distributions were modeled using GGD as described in equations 1.6 and 1.7 for every frame and geometric mean of shape parameter of all frames is computed to represent as a feature for that video.

2.1.3 Band Pass Statistics

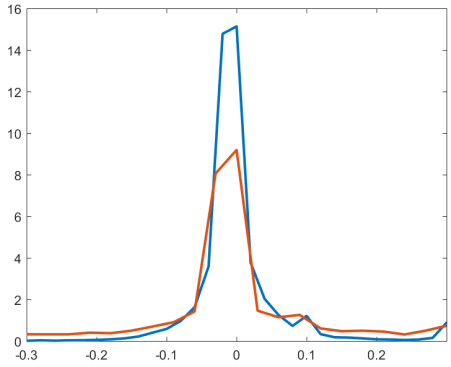
The frames of the optical flow field were subjected to band pass decomposition using steerable pyramids [15]. Here as well global statistics proved to be ineffective in capturing distortion information, hence local statistics were employed. Patch selection was done by using the same criteria discussed in the previous section. Figure 2.5 illustrates the histograms obtained from band pass statistics. The distinction between distorted and reference frame is more pronounced when compared with that of local velocity statistics.



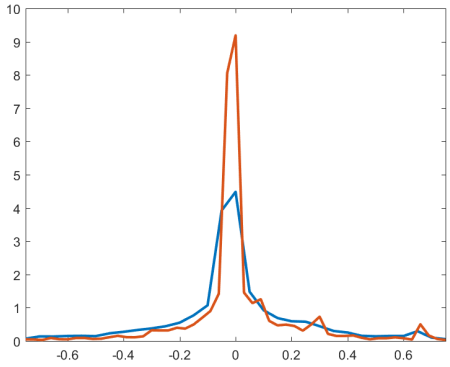
(a) Selected patches in reference frame



(b) Selected patches in distorted (wireless) frame

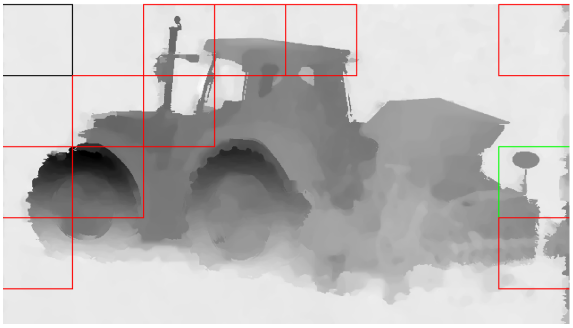


(c) Histogram of black border patch (blue - reference,orange - distorted)

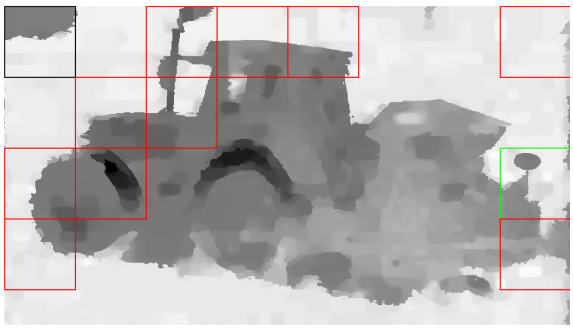


(d) Histogram of green border patch (blue - reference,orange - distorted)

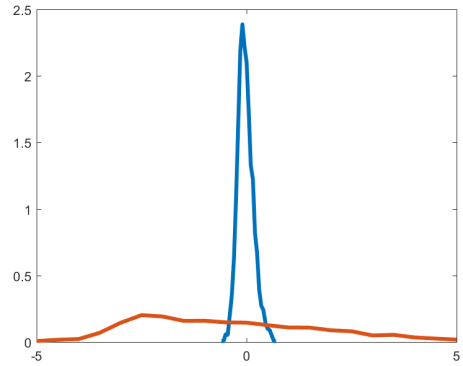
Figure 2.4: Local statistics of Horizontal field (10^{th} Video)



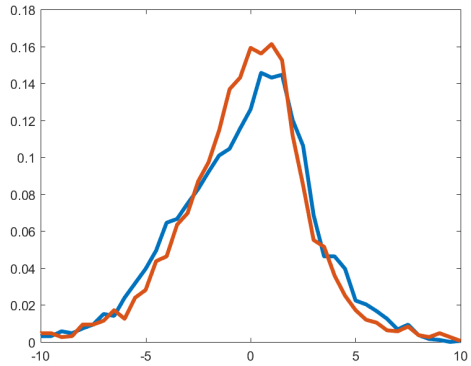
(a) Selected patches in reference frame



(b) Selected patches in distorted (wireless) frame



(c) Histogram of black border patch (blue - reference,orange - distorted, orientation - 0^0)

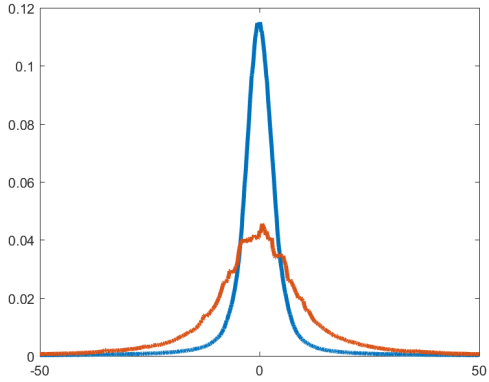


(d) Histogram of green border patch (blue - reference,orange - distorted, orientation - 180^0)

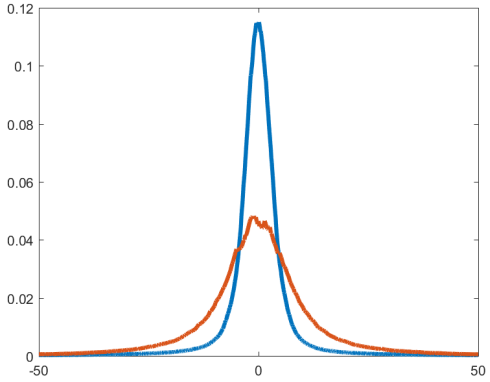
Figure 2.5: Local band pass statistics of Horizontal field (10^{th} Video)

2.2 Temporal Statistics

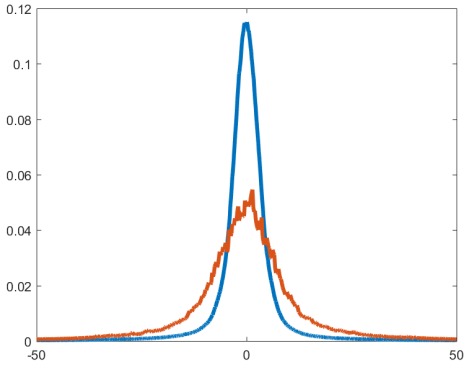
Spatial statistics only capture variations occurring in a single frame. However videos are inherently spatio-temporal. Video quality prediction algorithms generally employ frame differences to account for temporal artifacts. Figure 2.6 depicts the empirical histograms of Mean Subtracted (MS) coefficients of frame differences horizontal flow field. Similar observations were made in case of vertical field as well. The distributions were modeled using Generalized Gaussian Distribution (GGD) as shown in equations 1.6 and 1.7. Geometric mean is computed for all the shape parameters obtained from all frame differences of the video to represent the feature. Figure 2.2 iluustrates the variation of shape parameter γ over time for 3 different videos of varying quality. It can be inferred from the figure 2.2 that content plays a significant role in determining shape parameter. It is quite possible that videos having similar quality can have different shape parameter varying across time. This kind content-dependency poses a challenge in blind quality assessment as absolute values are less important than relative values.



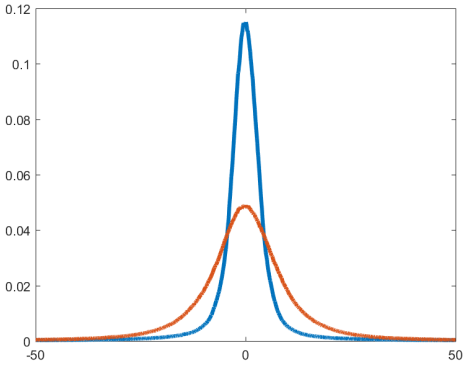
(a) Pristine and Wireless distortion



(b) Pristine and IP distortion



(c) Pristine and H.264 distortion



(d) Pristine and MPEG-2 distortion

Figure 2.6: Temporal statistics for various distortions for Horizontal flow field (10^{th} Video), blue - pristine, orange - distorted

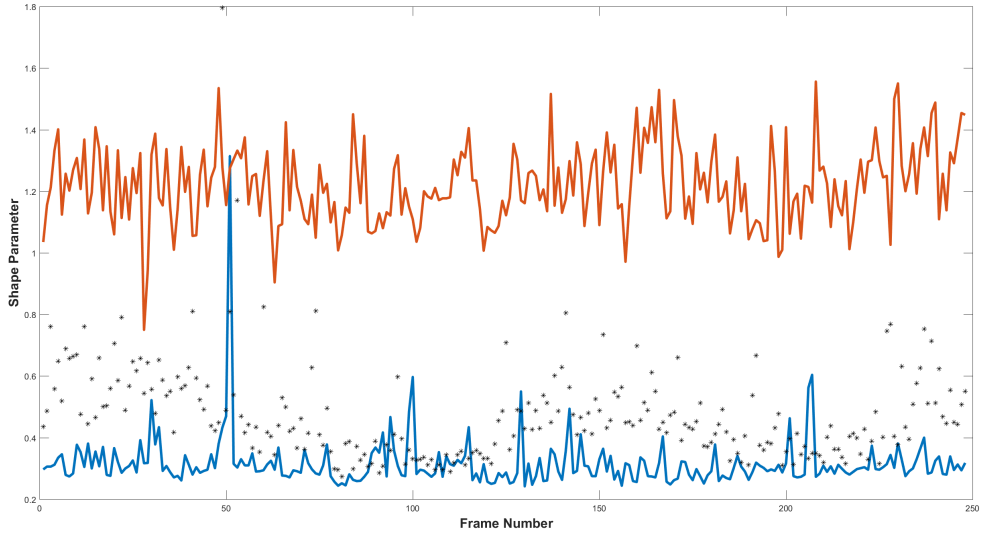


Figure 2.7: Temporal variation of γ. Orange - video having DMOS=39.19, Blue - low quality video with DMOS = 75.12,Black stars - pristine video with DMOS=0

2.3 Results

Table 2.1 shows Spearman Rank Order Correlation coefficients (SROCC) obtained across different combination of algorithms. The reported values are median SROCC values obtained by training SVR multiple times using different train and test combinations . Training data constituted 80% of the dataset and the rest were used for testing. Also train and test data had no content overlap across them. It is evident from the table that velocity statistics are hindering the performance, which characterizes the non-discrimanatory nature of features of spatial statistics. Temporal statistics tend to add no extra information as the performance does not improve significantly.

Algorithm	SROCC
BLIINDs	0.7081
BLIINDs+MSCN(Velocity)	0.6765
BLIINDs+Frame Difference	0.7072

Table 2.1: Comparison of algorithms

Algorithm	SROCC
BLIINDs	0.7081
MSCN(Velocity)	0.0042
Frame Difference	0.3286

Table 2.2: Independent contribution of the features

Table 2.2 represents the performance of features when evaluated in isolation with respect to other features. From the table it appears that velocity statistics are in fact uncorrelated with quality associated with that video. However table 2.3 suggests that

velocity statistics do perform well in case of wireless distortions. Table 2.3 also suggests that addition of temporal statistics of optical flow does not improve performance as it gives more or less similar results to that obtained by BLIINDs.

Algorithm	Wireless	IP	H.264	MPEG-2
BLIINDs	0.4286	0.4857	0.7143	0.8095
BLIINDs+MSCN(Velocity)	0.5	0.3714	0.7143	0.6905
BLIINDs+Frame Difference	0.4524	0.4857	0.6667	0.6905

Table 2.3: Performance across videos of specific distortions

2.4 Conclusion

In this project, employing optical flow statistics for blind video quality assessment was studied. The different methods experimented were built on the framework of video BLIINDs algorithm. For capturing spatial quality of every frame, BLIINDs employed NIQE index, however the same is inapplicable in case of optical flow as these frames of the flow fields aren't exactly natural images. The proposed metric based on MSCN coefficients for capturing spatial distortions wasn't particularly effective as observed in the results section. A more detailed study on designing features that effectively capture spatial distortions is required. With regard to temporal distortions, the distribution of frame differences were found to capture distortions, but during evaluation it was found that they didn't additively contribute towards performance enhancement. In other words their contribution was redundant when compared to the performance obtained using BLIINDs features. Therefore a more rigorous study on temporal distortions is needed in order to capture the variations observed in the flow fields that can possibly reflect on the video quality.

REFERENCES

[1] Andrew Burton and John Radford (1978). Thinking in Perspective: Critical Essays in the Study of Thought Processes. Routledge. ISBN 0-416-85840-6.

[2] Society of Robots, `http://www.societyofrobots.com/programming_computer_vision_tutorial_pt4.shtml`

[3] Optic Flow, Boston University `http://opticflow.bu.edu/`

[4] B. Horn and B. Schunck, Determining optical flow, Artificial Intelligence, vol. 17, pp. 185204, 1981.

[5] B. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in Proceedings of the 1981 DARPA Image Understanding Workshop, pp. 121130, 1981.

[6] Black, Michael J., and Paul Anandan. "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields." Computer vision and image understanding 63.1 (1996): 75-104.

[7] Brox, T., Bruhn, A., Papenberg, N. and Weickert, J., High accuracy optical flow estimation based on a theory of warping, European Conference on Computer Vision, 2004

[8] Fleet, D.J. and Jepson, A.D., Computation of component image velocity from local phase information, International Journal of Computer Vision, 1990

[9] Dosovitskiy, Alexey, et al. "Flownet: Learning optical flow with convolutional networks." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[10] E. Ilg, et al. "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks",IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul 2017.

[11] Sun, Deqing, Stefan Roth, and Michael J. Black. "Secrets of optical flow estimation and their principles." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010.

[12] Saad, Michele A., Alan C. Bovik, and Christophe Charrier. "Blind prediction of natural video quality." IEEE Transactions on Image Processing 23.3 (2014): 1352-1365.

[13] Roth, Stefan, and Michael J. Black. "On the spatial statistics of optical flow." International Journal of Computer Vision 74.1 (2007): 33-50.

[14] Mittal, Anish, Rajiv Soundararajan, and Alan C. Bovik. "Making a completely blind image quality analyzer." IEEE Signal Processing Letters 20.3 (2013): 209-212.

[15] Simoncelli, Eero P., and William T. Freeman. "The steerable pyramid: A flexible architecture for multi-scale derivative computation." Image Processing, 1995. Proceedings., International Conference on. Vol. 3. IEEE, 1995.

[16] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, KernlabAn S4 package for kernel methods in R, J. Statist. Softw., vol. 11, no. 9, pp. 120, Oct. 2004.

[17] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, Study of subjective and objective quality assessment of video, IEEE Trans. Image Process., vol. 19, no. 6, pp. 14271441, Jun. 2010.