U.Porto Journal of Engineering, 2:2 (2016) 14-25
ISSN 2183-6493

Received: 14 October, 2015
Accepted: 29 March, 2016
Published: 29 July, 2016

Multimedia Content Classification Metrics for Content Adaptation

Rui Fernandes[1], M. T. Andrade[2]
[1]Faculty of Engineering, University of Porto, Portugal (mpt09015@fe.up.pt); ESTIG-IPB, Bragança, Portugal; [2]INESC TEC, Faculty of Engineering, University of Porto, Portugal (mandrade@fe.up.pt)

Abstract
Multimedia content consumption is very popular nowadays. However, not every content can be consumed in its original format: the combination of content, transport and access networks, consumption device and usage environment characteristics may all pose restrictions to that purpose. One way to provide the best possible quality to the user is to adapt the content according to these restrictions as well as user preferences. This adaptation stage can be best executed if knowledge about the content is known *a-priori*. In order to provide this knowledge we classify the content based on metrics to define its temporal and spatial complexity. The temporal complexity classification is based on the Motion Vectors of the predictive encoded frames and on the difference between frames. The spatial complexity classification is based on different implementations of an edge detection algorithm and an image activity measure.

Subject Headings. Multimedia, Systems Interconnection
Author Keywords. Multimedia Classification, Temporal Complexity, Spatial Complexity, Multimedia Adaptation

1. Introduction

Today, there is a wide array of possibilities for consuming Multimedia content, from TV displays at home to portable devices on the go, with different types of transport and access networks, as depicted in Figure 1. Furthermore, there is already a great diversity of high definition (HD) content available and these consumption scenarios don't always meet the minimum requirements to enable an optimal HD content consumption. One way to overcome this limitation is to adapt the multimedia content, and content classification can provide insightful information to execute this task.

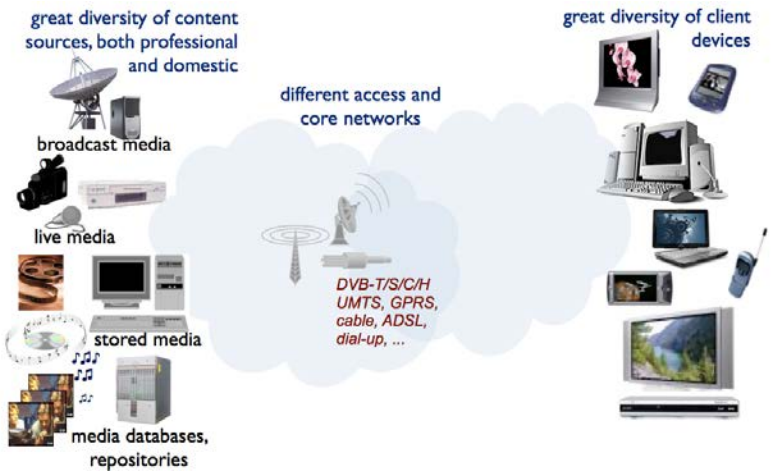


Figure 1: Present multimedia transmission situation. From (Andrade, 2009)

We present the implementation of different metrics to classify multimedia content, based on temporal and spatial complexity. This kind of classification can be very useful to help deciding the type of adaptation that should be performed on the content, to cope with existent restrictions imposed by the distribution chain and the consumption context. Depending on the type of content, which can be identified based on its spatial and temporal complexity, it may be advantageous to apply one specific adaptation over another one, so that the best possible subjective quality can be obtained whilst complying to the constraints. For example, if the network is imposing a limitation on the bandwidth available to transmit a video content, different forms of adaptation could eventually be applied to that content to reduce its bit rate below the constrained value (SNR adaptation, by manipulating the quantization step size; temporal adaptation, by reducing the frame rate; spatial adaptation, by reducing the spatial dimensions of the images; etc.). However, whilst complying to the imposed constraint, not all of them will deliver an adapted content presenting the same level of subjective quality. A content with high spatial detail and low temporal complexity may be wisely adapted by controlling the frame rate rather than the quantization step size or the spatial dimensions. Accordingly, knowing the complexity degree of the content can positively impact the adaptation decision, thus contributing to enable better users' experiences by offering always the best possible subjective quality under the imposed constraints.

2. Temporal Complexity Metrics

There are several possibilities available to quantify the temporal complexity of a multimedia content. The next subsections present the implemented approaches to generate this type of classification.

2.1. Intensity of Motion

One way to classify the temporal complexity of a sequence is to look into the magnitude of the motion vectors (MVs) computed for the predicted frames of the sequence. In (Amel and Abdessalem and Abdellatif, 2010), the authors present an algorithm to measure the intensity of motion on a frame basis, using a five level classification scale.

Their algorithm determines initially the spatial activity matrix

$$C_{mv} = \{R(i,j)\} \tag{1}$$

in which a single matrix element is defined by

$$R_{xy}(i,j) = \sqrt{x(i,j)^2 + y(i,j)^2} \tag{2}$$

In these equations, (i,j) define the block indices and $x(i,j)$ and $y(i,j)$ represent, respectively, the horizontal and vertical motion vectors of the corresponding block.

The temporal complexity predictor used is the standard deviation value of the motion vectors present in the frame, thus, the authors compute the mean average value of the activity matrix

$$C_{mv}^{avg} = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} C_{mv}(i,j) \tag{3}$$

which is used afterwards to compute the intensity of motion

$$\sigma_{fr}^2 = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(C_{mv}(i,j) - C_{mv}^{avg}\right)^2 \tag{4}$$

Here, M and N define the number of horizontal and vertical macroblocks present in the frame.

The temporal complexity classification of the frame is obtained by taking into consideration the levels defined in Table 1.

Classification	Dynamics of the standard deviation of motion vectors, σ
1 – Very Low Intensity	0 ≤ σ ≤ 3.9
2 – Low Intensity	3.9 ≤ σ ≤ 10.7
3 – Medium Intensity	10.7 ≤ σ ≤ 17.1
4 – High Intensity	17.1 ≤ σ ≤ 32
5 – Very High Intensity	σ ≥ 32

Table 1: Temporal complexity classification scale.
Adapted from (Amel and Abdessalem and Abdellatif, 2010)

Equations 3 and 4 are based on the assumption that there is only one motion vector per macroblock, which may not be true for the type of encoding used on our simulations, which is the H.264 encoder. Thus, our implementation of this methodology takes this fact into consideration and the needed changes were introduced to account all motion vectors, including the multiple motion vectors per macroblock that may appear in the encoded frame.

We have implemented this metric to establish the original (1) frame-by-frame classification, and also extended it to perform (2) Group Of Pictures (GOP)-by-GOP or group of GOPs-by-group of GOPs classification, (3) scene-by-scene classification and (4) multimedia clip classification.

Two different approaches were tested for scenarios (2), (3) and (4):

1. The average intensity of motion, of all frames under consideration, is determined and used along with the original classification scale, established in Table 1, to define the classification;
2. The statistical mode of the frames classifications is determined and attributed as the final classification.

The scene-by-scene classification needs to establish the boundaries within the analysis has to be performed, which means it needs to execute scene change detection. The used algorithm to achieve this objective is based on the difference between the luminance component of two frames, $z_p^l(i)$, and was chosen because it provides a good relation between effective scene detection efficiency and computational implementation cost. The algorithm performing the difference determination is based on Equation 5 (DEEC/UC).

$$z_p^l(i) = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} |f_i(x,y) - f_{i+l}(x,y)| \quad (5)$$

Consecutive frames are used in our implementation, thus, $l = 1$. The index p refers to the fact that it's a pixel based operation.

If $z_p^l(i) > k_p$ frame i delimits a scene. k_p, the scene detection threshold was determined experimentally and was defined as 30.

After having knowledge about the scenes boundaries, the classification is executed using the exact same methodology previously presented for the GOP classification.

Finally, using this method and considering the existence of N frames and N_I intra frames, regardless of the scenario under analysis, the values determined for the individual frames can be used to generate a set of characteristics about the multimedia content, namely the:

- mean average of the mean averages of the motion magnitude of the analyzed frames, $\overline{C_{mv}^{avg}} = \frac{1}{N-N_I}\sum_{i=1}^{N-N_I} C_{mv\,i}^{avg}$
- standard deviation of the individual analyzed frames intensity of motion, $\sigma_{clip} = \sqrt{\frac{1}{N-N_I}\sum_{i=1}^{N-N_I}\left(\sigma_i - \overline{C_{mv\,i}^{avg}}\right)^2}$.

These characteristics allow the generation of more knowledge about the multimedia clip in terms of its motion complexity degree and correspondent variation throughout the sequence. Lower values of σ_{clip} identify the same type of temporal complexity throughout the sequence and higher values identify different types of temporal complexities present throughout the sequence.

2.2. Temporal Index

This metric is based on the determination and observation of the variation of the Temporal perceptual Information (TI) defined by the ITU to establish the temporal complexity (*ITU-T Study Group 12, 1999*). Our implementation was conceived considering the modifications proposed in (Korhonen and Reiter and Ukhanova, 2013), to the index definition, to avoid overemphasizing the impact of high motion or high spatial details of temporarily appearing objects, resulting in Equation 6.

$$TI = mean_{time}\left(std_{space}[F_n - F_{n-1}]\right) \quad (6)$$

From Equation 6, the index consists on the mean average over time of the standard deviation values of the difference between two consecutive frames of the multimedia content.

In order to have current information available, the mean average is updated for every new processed frame. This also allows the availability of correct information when the content is divided in segments, where different adaptation decisions can/must be made.

2.3. Temporal Uniformity Index

Since the TI metric was very generic, a new complementary metric was defined in (Korhonen and Reiter and Ukhanova, 2013), enabling a more accurate characterization of the multimedia content by performing uniformization. This metric is also a modification of the definition established in (*ITU-T Study Group 12, 1999*) and divides the frame into blocks, determining the temporal index for each individual block, by implementing the operations depicted in Equation 7 blockwise.

$$T_{n,m} = std_{space}\left(B_{n,m} - B_{n-1,m}\right) \quad (7)$$

In Equation 7, n represents the frame number and m represents the block in the frame. It is assumed that there are M blocks per frame and that the block size is predefined before implementation.

The standard deviation of the $T_{n,m}$ values of all blocks in a frame is determined and uniformed by the mean average of the $T_{n,m}$ block values. The temporal uniformity index is computed by taking the mean over time the resulting values, per frame, as presented in Equation 8:

$$TUI = mean_{time}\left(std_{space}\left[T_{n,1\ldots M}\right]/mean_{space}\left[T_{n,1\ldots M}\right]\right) \quad (8)$$

This metric can be combined with the TI metric to achieve a better temporal complexity categorization of the multimedia content, as presented in (Korhonen and Reiter and Ukhanova, 2013), according with Table 2.

	TUI low	*TUI* high
TI low	Static, low motion scene	Mostly static, some moving objects
TI high	Overall motion: panning, ripple, etc.	Overall motion with varying intensity

Table 2: Qualitative classification based on TI and TUI indexes. Adapted from (Korhonen and Reiter and Ukhanova, 2013)

3. Spatial Complexity Metrics

Just as with the temporal complexity, spatial complexity can be derived using different approaches. The following subsections present the implemented methodologies for the spatial complexity determination.

3.1. Sobel Filtering

Spatial complexity is usually established through the use of edge detection algorithms over the luminance component of the sequence. A variation of the Sobel edge detection algorithm was selected and computed according to the definition established in (*ITU-T Study Group 12, 1999*). Consequently, one of the methodologies defined in (Yu and Winkler, 2013) was employed to obtain a measure of the spatial complexity of a frame. This method uses the resulting filtered images using the horizontal and vertical Sobel kernels to compute the spatial information through the Pythagoras' theorem, as defined in Equation 9.

$$SI_r = \sqrt{s_h^2 + s_v^2} \tag{9}$$

The spatial index complexity is then determined by taking the mean of the spatial information values for all pixels of the frame (P): Equation 10.

$$SI_{mean} = \frac{1}{P}\sum SI_r \tag{10}$$

This method was chosen because it provides good results at low complexity for its implementation.

3.2. Image Activity Measure

To assess the quality of the obtained results, a second metric was used, based on both vertical and horizontal local gradients calculated over the luminance component of the frames and presented in Equation 11 (*Engelke et al, 2009*).

$$\widetilde{f_4} = \frac{1}{MN}\left(\sum_{i=1}^{M-1}\sum_{j=1}^{N}|I(i,j) - I(i+1,j)| + \sum_{i=1}^{M}\sum_{j=1}^{N-1}|I(i,j) - I(i,j+1)|\right) \tag{11}$$

M and N, in Equation 11, represent the content dimensions.

Our results confirmed that this metric, presented in (*Engelke et al, 2009*), quantifies image activity very accurately, delivering higher values as the spatial complexity increases and is in accordance with the results of the Sobel edge detection technique.

3.3. Spatial Index

A spatial complexity metric, analogous to the temporal complexity metric referred in 2.2., was implemented to quantify the spatial complexity of the multimedia content. Its definition, as

devised in (Korhonen and Reiter and Ukhanova, 2013), consists in a modification of the original definition presented in (*ITU-T Study Group 12, 1999*), resulting in Equation 12:

$$SI = mean_{time}\left[mean_{space}Sobel(F_n)\right] \quad (12)$$

From Equation 12, the index computes the Sobel filtering operation of a frame, executing the mean average value of the filtering outputs to return a value per frame. Afterwards, a mean average of these final frame values is computed to generate the spatial index. This mean average over time allows to establish the spatial complexity classification over a series of frames instead of characterizing a single frame.

In order to have a current value, the mean average is updated every time a new frame is processed.

Higher index values indicate higher spatial complexity per group of analyzed frames.

3.4. Spatial Uniformity Index

Analogously to the temporal metric implementation discussed on 2.3., a spatial complexity metric, in which uniformization is performed, was also implemented. This metric, accordingly with Equation 13, divides the frames in blocks, applies Sobel filtering to each individual block and mean averages the Sobel filtering outputs to generate a value per block.

$$S_{n,m} = mean_{space}\left(Sobel\left(F_{n,m}\right)\right) \quad (13)$$

In Equation 13, n represents the frame number and m represents the block in the frame. It is assumed that there are M blocks per frame and that the block size is predefined before implementation.

The standard deviation of these values is determined and the result is normalized by the mean average of the final block values. The mean average of the resulting values per frame is computed to create the spatial complexity metric, SUI, as shown in Equation 14.

$$SUI = mean_{time}\left(std_{space}\left[S_{n,1\ldots M}\right]/mean_{space}\left[S_{n,1\ldots M}\right]\right) \quad (14)$$

As previously referred, the final mean average value is updated every time a new frame is processed and higher index values establish higher spatial complexity per group of analyzed frames.

The metrics SI and SUI can also be combined, similarly to the combination previously established between TI and TUI. The combination between the metrics allows a better spatial complexity classification of the multimedia content, as presented in (Korhonen and Reiter and Ukhanova, 2013), according with Table 3.

	SUI low	SUI high
SI low	Lot of smooth and uniform surfaces	Mostly smooth, some detailed objects
SI high	Lot of detailed patterns	Mixture of more and less detailed patterns

Table 3: Qualitative classification based on SI and SUI indexes. Adapted from (Korhonen and Reiter and Ukhanova, 2013)

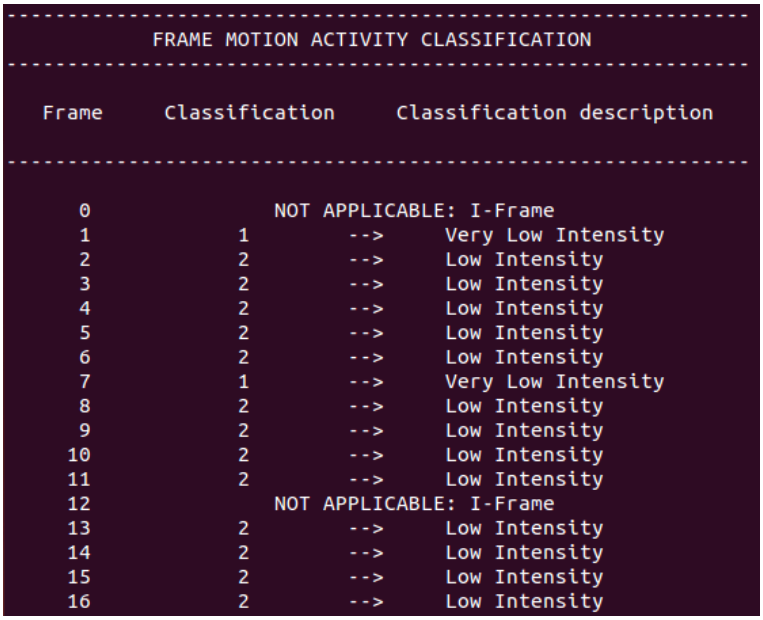
4. Results

The implemented metrics deliver a classification of the content, regarding its temporal and spatial complexity, which is subsequently used by the adaptation decision algorithm. This way, this module is enriched with meaningful information to choose the optimal adaptation to be executed for the present situation. Some of the obtained results, for each implemented metric, are presented in the next subsections.

4.1. Intensity of Motion

Figure 2 shows the obtained results for the implementation of the metric presented in (Amel and Abdessalem and Abdellatif, 2010), for classifying individual frames (scenario (1)). The first column identifies the frame under analysis and the other two columns present the computed classification for that frame.

As it can also be seen in Figure 2, this metric cannot be executed over Intra frames, since this type of frames does not have motion vectors associated with it to be analyzed.

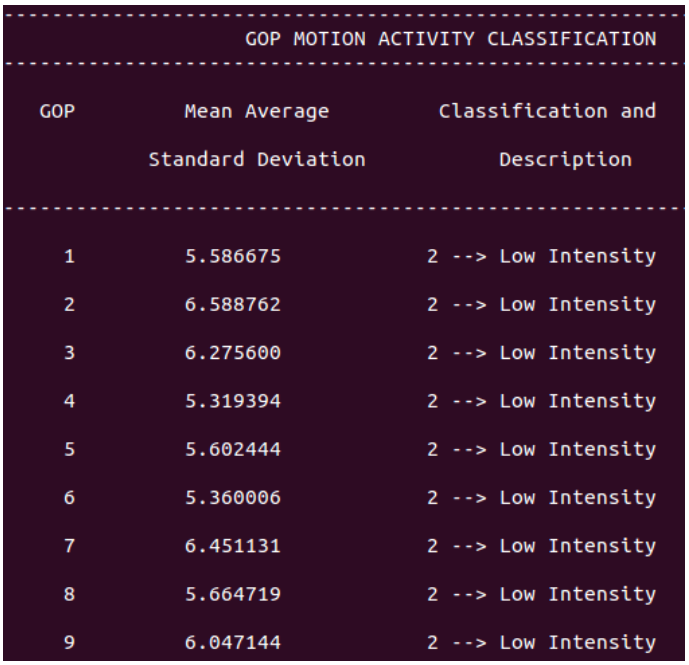


FRAME MOTION ACTIVITY CLASSIFICATION

Frame	Classification	Classification description
0	NOT APPLICABLE: I-Frame	
1	1 -->	Very Low Intensity
2	2 -->	Low Intensity
3	2 -->	Low Intensity
4	2 -->	Low Intensity
5	2 -->	Low Intensity
6	2 -->	Low Intensity
7	1 -->	Very Low Intensity
8	2 -->	Low Intensity
9	2 -->	Low Intensity
10	2 -->	Low Intensity
11	2 -->	Low Intensity
12	NOT APPLICABLE: I-Frame	
13	2 -->	Low Intensity
14	2 -->	Low Intensity
15	2 -->	Low Intensity
16	2 -->	Low Intensity

Figure 2: Temporal complexity classification of individual frames of a multimedia sequence. Analysis shown for sequence office.h264

Figure 3 depicts the obtained results for GOP classification under the first implementation approach. The first column identifies the GOP under analysis, the second column provides the computed mean average of the intensity of motion of all frames of the GOP and the final column presents the obtained classification based on the second column values and the scale presented in Table 1.



GOP MOTION ACTIVITY CLASSIFICATION

GOP	Mean Average Standard Deviation	Classification and Description
1	5.586675	2 --> Low Intensity
2	6.588762	2 --> Low Intensity
3	6.275600	2 --> Low Intensity
4	5.319394	2 --> Low Intensity
5	5.602444	2 --> Low Intensity
6	5.360006	2 --> Low Intensity
7	6.451131	2 --> Low Intensity
8	5.664719	2 --> Low Intensity
9	6.047144	2 --> Low Intensity

Figure 3: Temporal complexity classification of individual frames of a multimedia sequence. Analysis shown for sequence office.h264

The usage of the second approach provides two different cases. The first case has similar classifications to the ones obtained in the first approach. The second has classifications that may differ from the ones obtained in the first approach.

GOP		1
GOP limits		Frames 0 and 11
Mean average standard deviation		5.59
Mean average standard deviation classification		2 → Low Intensity
Number of frames of the GOP with classification:	1	2
	2	9
	3	0
	4	0
	5	0
Statistical mode classification		2 → Low Intensity

Table 4: GOP temporal complexity classification from the two defined approaches with similar results

Table 4 presents the case of similitude of classifications generated by both approaches. This similarity occurs whenever the type of movement is maintained throughout all frames of the GOP. In this situation the numerical analysis will be equal to the statistical mode analysis.

Whenever there exists more variability throughout the GOP, regarding the type of movement, the classifications of the two approaches may end up being different since the classification scale is not linear. This scenario is presented in Table 5, for a GOP that mixes low and high intensity movement.

GOP		30
GOP limits		Frames 348 and 359
Mean average std		15.83
Mean average standard deviation classification		3 → Medium Intensity
Number of frames of the GOP with classification:	1	2
	2	3
	3	0
	4	6
	5	0
Statistical mode classification		4 → High Intensity

Table 5: GOP temporal complexity classification from the two defined approaches with different results

As it can be seen, the classifications are different and, in this case, the statistical mode approach disregards the fact that almost half of the frames have low complexity whilst the first approach considers every element on its calculations. For that reason, the first approach provides more accurate results and its use should be preferable relatively to the second approach, in this type of scenario.

Figure 4 presents an example for the determination of the set of extra characteristics. The 6.21 value generates the clip classification of Low Intensity. The 1.67 value, being a low value, confirms that this classification is present throughout the multimedia clip.

```
------------------------------------------------------------
            FRAME MOTION ACTIVITY CLASSIFICATION
              standard deviation of the motion
                magnitude standard deviation
              of the multimedia content frames
------------------------------------------------------------

  Content       Value     Motion magnitude     Standard deviation
                                mean average          mean average

 office.h264     1.67           8.16                  6.21
```

Figure 4: Extra characteristics determination

4.2. TI and TUI

The results for these metrics are presented for the combination of the two, as mentioned in 2.3. To better understand the classification provided by these metrics, Figure 5 presents the final results for contents distributed through the four different zones identified by the metrics combination.

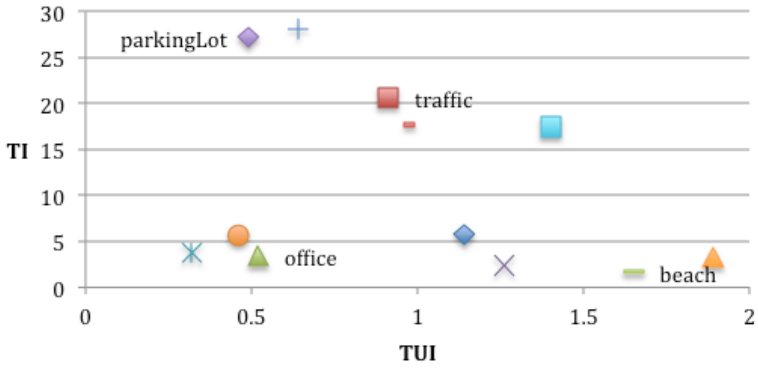


Figure 5: TI and TUI based temporal complexity classification

This classification has the advantage of being more contextually complete when compared with the other implemented metrics for the temporal complexity classification. It classifies the content through four different contextual groups, as defined in Table 2.

4.3. Sobel Filtering and Image Activity Measure

The obtained results with the implementation of these metrics are consistent between themselves and provide the same type of value variation for different spatial complexity sequences. When one increases so does the other and the same happens for the opposite case. This can be seen in Figure 6.

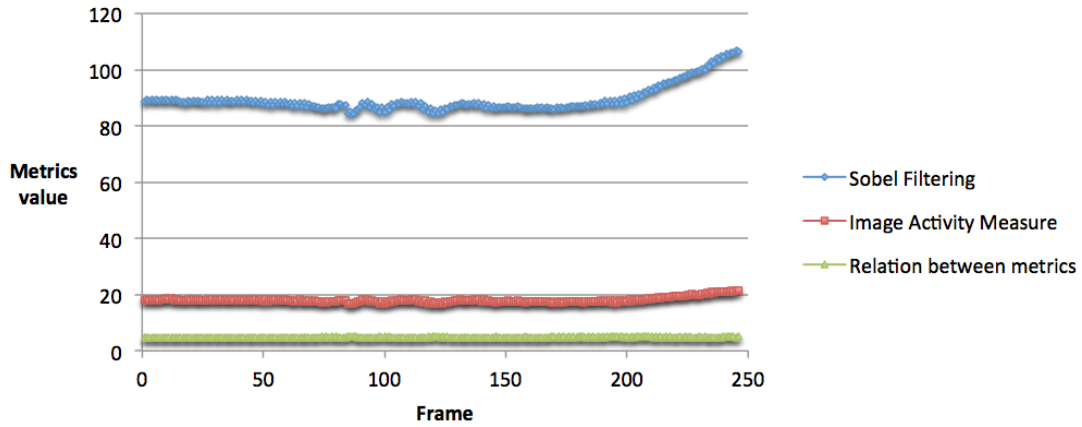


Figure 6: Sobel Filtering and Image Activity Measure values for spatial complexity identification: office.h264 sequence: approximately 10 seconds

Figure 7 presents the same type of results, for these metrics, to another sequence.

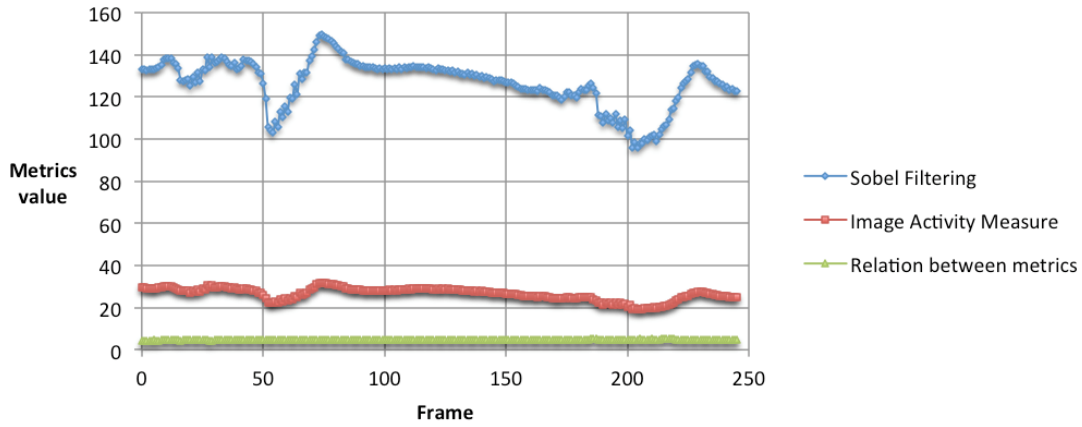


Figure 7: Sobel Filtering and Image Activity Measure values for spatial complexity identification: traffic.h264 sequence: approximately 10 seconds

From the analysis of Figure 8, it can be seen that the relation between the metrics depends on the content, however it is almost linear, regardless of the content. Since they provide the same type of information, in different ranges, the choice between these metrics can then be executed considering their computational implementation costs.

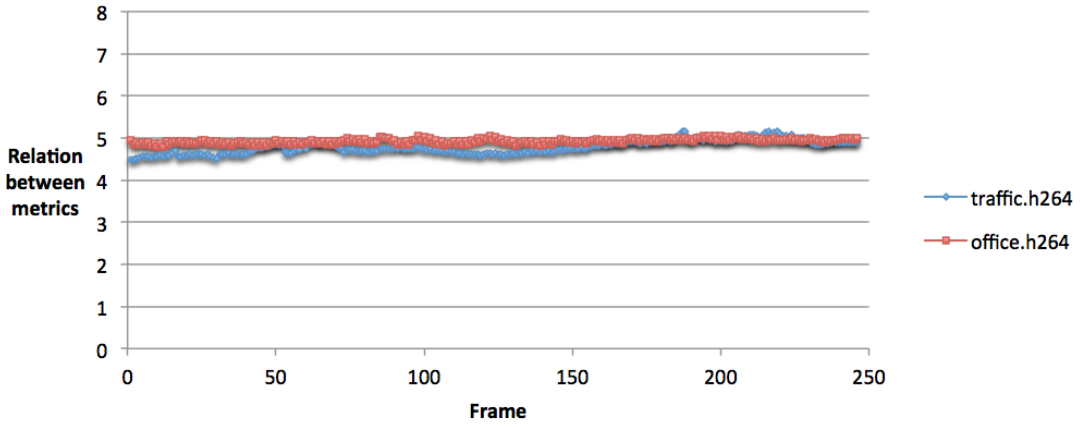


Figure 8: Relation between metrics for different contents: approximately 10 seconds

4.4. SI and SUI

Just as with TI and TUI metrics, the results for the SI and SUI metrics are presented coupled together, to take advantage of the better contextually classification of contents presented in Table 3. Accordingly, this combination of SI with SUI defines the content classification within the four available possibilities, as represented in Table 3.

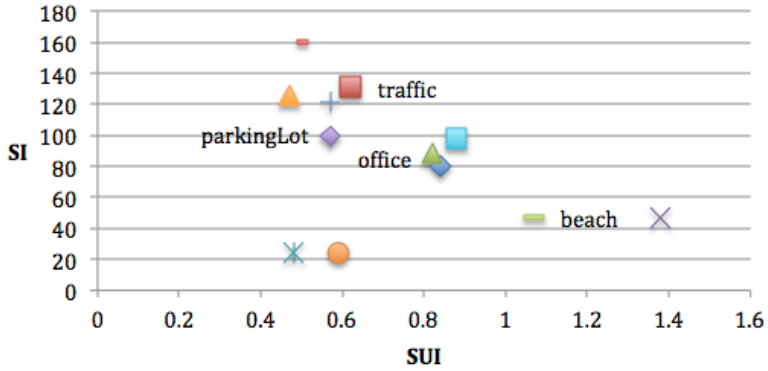


Figure 9: SI and SUI based spatial complexity classification

In Figure 9, each symbol represents a different content, and the Figure depicts contents for all four possible classifications.

5. Conclusions

The implemented metrics proved to be adequate to classify multimedia content by measuring with a good level of accuracy its temporal and spatial complexity on a frame, GOP, scene or complete clip basis. This classification is crucial for an adaptation decision engine to be able to select the best adaptation among several possible ones aiming at offering the best possible Quality of Experience to the user.

By comparing the automatic classification of the content using the temporal complexity metrics with a visual inspection of the content, it is possible to conclude that it delivers reasonably accurate results. Still, the results reliability of the statistical mode strategy, devised

for groups of frames classification, is directly interconnected with the type of content under classification. Furthermore, an analysis using derivatives is being conducted to assess whether it is possible to perform the classification without the need to process all frames. The results for the spatial complexity metrics, Sobel filtering and Image Activity Measure, yield similar classifications with an almost linear relation between the metrics' values in which the classification is based.

Our results show that, among the implemented metrics, the combination of TI with TUI and SI with SUI provide, among the implemented metrics, the best contextual characterization of the context, at the expense of a small increase in computational implementation cost. Moreover, they represent analogous analysis to define the spatial and temporal complexities and, when combined, generate 16 different possible classes of contents, which is more than enough. For these reasons these are the metrics chosen to classify the multimedia contents under transmission, which may require adaptation. This means that the resulting classification of these metrics is to be passed to the adaptation decision engine, so that this engine can use this information in the decision-taking process.

References

Andrade, Maria Teresa. 2009. "A look at the new multimedia landscape". Presentation at MAP-Tele Doctoral Course 2009-2010. Accessed July 18, 2016. https://paginas.fe.up.pt/~mandrade/presentations/mapTele09.pdf.

Amel, Abdelati Malek, Abdessalem, Ben Abdelali and Abdellatif, Mtibaa. 2010. "Video shot boundary detection using motion activity descriptor". *Journal of Telecommunications*, 2 (1):54-59.

DEEC/UC. "Relatório Técnico Final". Projeto AdNOW, Relatótio no. 11499. DEEC/UC.

ITU-T Study Group 12. 1999. "Subjective video quality assessment methods for multimedia applications". ITU-T Recommendation P.910, International Telecommunication Union.

Korhonen, J., Reiter, U. and Ukhanova, A. 2013. "Frame rate versus spatial quality: Which video characteristics do matter?". *Visual Communications and Image Processing 2013*, 1-6. Accessed July 18, 2016. DOI: 10.1109/VCIP.2013.6706381.

Yu, Honghai and Winkler, Stefan. 2013. "Image complexity and spatial information". *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, 12–17. Accessed July 18, 2016. DOI: 10.1109/QoMEX.2013.6603194.

Engelke, Ulrich, Kusuma, Maulana, Zepernick, Hans-Jürgen and Caldera, Manora. 2009. "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging". Signal Processing: Image Communication, 24 (7):525-547. Accessed July 18, 2016. DOI: 10.1016/j.image.2009.06.005.