# A Quantitative Method for Graphical Representation of H.264 Video Content Dynamics

Michail-Alexandros Kourtis[1], Harilaos Koumaras[2], Charalampos Skianis[1], Fotios Lazarakis[2]

[1]Dep. of Information & Com. Systems Engineering
University of the Aegean
Karlovasi, Samos, Greece
{akiskourtis, cskianis}@aegean.gr

[2]Institute of Informatics and Telecommunications
National Centre of Scientific Research "Demokritos"
Athens, Greece
{koumaras, flaz}@iit.demokritos.gr

***Abstract:*** **This paper presents a method for graphical representation of H.264 video content dynamics, based on the extraction of motion information from the video signal in order to define the content's spatiotemporal dynamics levels. For this purpose, the paper introduces a motion compensation and content residual grid, where each video signal is represented as a point with coordinates that are defined by its motion dynamics and content residual, respectively. The proposed method can be effectively integrated as part of the decision process in video adaptation systems, which perform spatial or temporal (or both) adaptation actions, like MPEG-DASH, in order to optimize the video service delivery.**

***Keywords- Video Graphical Representation, Motion Vector Analysis, Video Dynamics, H.264.***

## I. INTRODUCTION

In recent years, due to the ever-growing demand for multimedia services, a vast and increasing request for video content services has been observed. This has created a substantial need for high-quality, trust-worthy and efficient video services. However, this proved to be a complex and difficult task, due to the fact that the video stream is usually transmitted over error-prone IP network environments, whose network impairments (e.g. packet loss, delay and delay variation for IP networks) may cause quality degradation. In order to meet the high-end user demands, and simultaneously tackle effectively the obstacles created by the heterogeneous network environment, several solutions have been proposed.

One of the proposed solutions is the video adaptation framework using the MPEG standard Dynamic Adaptive Streaming over HTTP (MPEG-DASH) [1, 2]. Depending on network conditions or user feedback, the video delivery stream is dynamically switched seamlessly among different pre-encoded streams, which have been adapted either in the spatial (i.e. resolution) or the temporal (i.e. frame rate) domain.

For the decision making process of MPEG-DASH and for the successful preparation of the video signal at various spatial or temporal adapted versions, it is crucial to differentiate the spatiotemporal video content dynamics of the test signal. For example, video content with high temporal and low spatial dynamics will be significantly degraded by a temporal adaptation decision action (introducing jerky motion and pauses), thus spatial adaptation should be preferred instead. Conversely, video signals with high spatial and low temporal dynamics, should be adapted by applying a temporal adaptation process in order to minimize the perceptual impact of the adaptation action, and vice versa.

Therefore, it is necessary for the optimization of the decision making process of adaptive systems, like MPEG-DASH, to consider in the adapted stream preparation/encoding process, appropriate methods that differentiate the video content dynamics of the signal both at the spatial and temporal domain.

This paper proposes a method to quantify and depict the spatiotemporal video content dynamics of each video signal under test onto a spatiotemporal plane. More specifically, a mapping of the spatial and temporal domains on a motion compensation and content residual grid is proposed by utilizing the signal's motion vectors. Towards this mapping, two appropriate metrics are proposed: The content residual metric and the motion compensation metric.

The content residual metric refers to the video's average motion vector length, thus the activity level of the video's elements, meaning the shift between successive frames. The motion compensation metric measures the average motion vector number of the video, thus depicting the overall motion activity of a video sequence, how many macroblocks changed position during the video sequence.

Following this introductory section, the rest of the paper is organised as follows: Section II proposes a graphical representation of the spatiotemporal dynamics, while section III and IV discusses how the analysis of the motion vectors can be exploited for the scope of the paper and two appropriate metrics are presented. Section V provides experimental results of the proposed methods for evaluation and validating purposes, while Section VI concludes the paper.

## II. SPATIOTEMPORAL PLANE

The impact of either spatial or temporal adaptation on a video signal may have a different impact on the perceptual quality of the video sequence depending on the spatiotemporal dynamics of the content. It is well established from the fundamental principles of video coding theory that action clips with high dynamic content are perceived as more degraded in

comparison to sequences with slow-moving clips, considering identical encodings.

In order to differentiate a video clip according to the spatial and temporal complexity of its content, a spatiotemporal plane [3] is considered as depicted on Figure 1. According to this approach, each video clip can be classified to four categories depending on its spatiotemporal content dynamics.

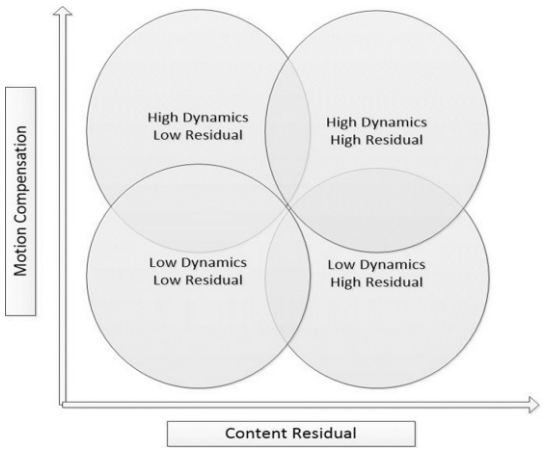


Figure. 1. Spatiotemporal Plane

For the graphical representation of each test signal on the spatiotemporal plane according to its content dynamics, two discrete metrics (one for the spatial and one for the temporal domain) are introduced in this paper, which are based on the data analysis of the motion vectors of each video signal frame. The proposed metrics can be applied on every video signal which has been encoded according to a codec that utilizes motion vectors in the motion estimation process. However, for the experimental needs and the validation of the proposed method, this paper presents experimental results of video signals that have been encoded by H.264/MPEG-4 AVC [4]. Thus, for the purposes of these experiments, standardized H.264/AVC bit streams of each video signal were used in order to extract the motion vector information.

## III. Motion Vector Analysis

In the H.264/AVC video coding standard each frame is encoded using one out of the three different prediction levels: Intra-frame (I), Inter-frame Predictive (P) and Inter-frame Bidirectional predictive (B). Each frame consists of 16x16 sized regions called macroblocks, which are further subdivided into transform blocks and may be further subdivided into prediction blocks of various sizes [4].

During the prediction process, the encoding algorithm uses motion estimation to search for an area (macroblock) in the reference frame(s) (past, future or both) in order to find a corresponding matching region. The result of this process is translated into a motion vector for the region under-encoding, which denotes the position where the region will move. Thus, during motion estimation the best matches between reference and current frames are detected and this match is further improved by motion compensation, which calculates the residuals of the motion estimated frame and the actual frame. So, adding this motion compensated residual information on the motion estimated frame, an accurate and efficient prediction of the current frame can be performed, using regions of past or future frames. This data is formatted and stored in the encoded H.264/AVC bit stream file on the corresponding macroblocks.

In order to extract the motion vector information for each frame from the encoded video bit stream file, a motion vector extraction module was implemented and added to the standard H.264/AVC decoder [4]. This additional function stores the motion vector information in the following XML-formatted manner to a file, for later use:

```
<Frame id="integer">
      <Macroblock id="integer">
            <MotionVector x="int"y="y"/>
      </Macroblock>
      .
      .
      .
      <Macroblock id="integer">
            <MotionVector x="int"y="y"/>
      </Macroblock>
</Frame>
```

As can be observed, the information stored in this file is organized by frame on the first level and by macroblock on the second level. In each macroblock section resides its motion information in the form of the motion vector's x and y coordinates.

## IV. Proposed Spatiotemporal Metrics

Following, the extraction of the motion vectors data and the representation of them by an appropriate XML schema, the application of the proposed two metrics (i.e. the content residual metric and the motion compensation metric) is performed for the appropriate mapping of the test signal on the spatiotemporal plane.

### *A.* Content *Residual* Metric

For every motion vector $v_i$ of a video frame, its norm is calculated by the equation (1) using its $x$ and $y$ coordinates:

$$|\vec{v_i}| = \sqrt{x^2 + y^2} \qquad (1)$$

Afterwards for each frame, the total sum of all its motion vector norms $S_f$ is calculated:

$$S_f = \sum_{i=1}^{n} |\vec{v_i}| \qquad (2)$$

Where $n$ is the total number of motion vectors contained in the corresponding frame. Subsequently we calculate the average of all $S_f$ for the whole video signal duration in equation (3):

$$AvgLengths_v = \frac{\sum_{i=1}^{k} S_{f_k}}{k} \qquad (3)$$

Where $k$ is the number of the total frames of the video signal. Thus, the result of this equation is the average motion vector norm of the video signal, which is proposed as the content

residual metric for the quantification of the video's total pixel and artifact shift from frame to frame.
Below on figure 2, we present a visualized example of a video frame with a high valued residual index content. Notice the high valued motion vectors of the fast moving ball on the right frame.

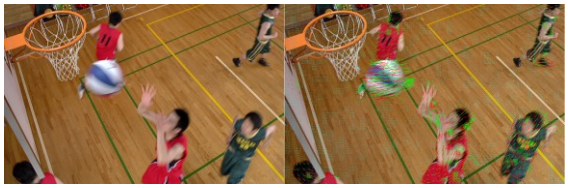

Figure. 2. A video frame with a high valued residual index content

### B. *Motion Compensation Metric*

Since a high content diversity among the depicted video signal's objects will have an impact on the number of motion vectors per frame, then the proposed metric for the quantification of the spatial aspect of the video content dynamics is proposed the average number of motion vectors per frame, which is calculated in equation (4):

$$AvgTotals_v = \frac{\sum_{i=1}^{k} t_i}{k} \qquad (4)$$

Where $t_i$ is the total number of motion vectors per frame, and $k$ the total number of frames of the video signal under test. The average motion vector total per frame creates the motion compensation metric. This is derived by the fact that an index of the video's averaged total motion activity can give us the video's motion compensation.

Below on figure 3 we present a visualized example of a video with a high valued motion compensation index. Notice the differences depicted on the residual frame on the right.

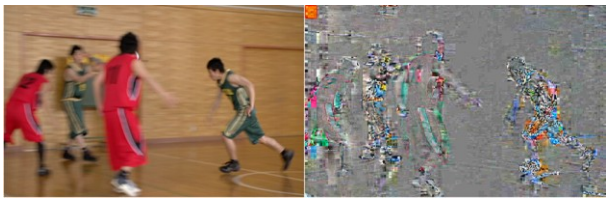

Figure. 3. A highly motion compensated video frame.

## V. Experimental Results and Method Validation

The proposed graphical representation method is based on the quantitative calculation of the two proposed metrics for the measurement of the motion compensation and content residual aspect of the video's content dynamics.

Considering the values of the two metrics as coordinates on the proposed content residual and motion compensation grid of figure 1, each video sequence (subject to short duration and homogeneous content) is depicted as a spot in the proposed spatiotemporal plane, where the vertical axis refers to the motion compensation aspect of the content dynamics through the proposed metric (4) (i.e. the mean motion vector number) and the horizontal axis refers to the content residual aspect of the content dynamics through the proposed metric (3) (i.e. the mean motion vector norm).

The advantages of the proposed two dimensional representation of the spatiotemporal video content dynamics are the unique graphical characterization of each test signal and the relative comparison of content residual and motion compensation dynamics, while the main drawback is the difficulty in the characterization of highly heterogeneous test signals.

For the experimental needs of this report towards the validation of the proposed method, 16 reference video clips were used [5], which are representative of various spatiotemporal video content dynamics. The test signals have spatial resolution 416x240, 832x480, 1280x720 and 1920x1080. The videos and their corresponding resolution are listed in Table I, along with a representative video frame on Figure 4.

Figure. 4. Snapshots of every video under test.

For the validation needs of this report all videos were encoded from their original YUV format to ISO AVC High Profile with quantization parameter equal to 12. Across the encoding process the reference software was used. During the decoding process the decoder reference software was modified in order to export an XML file with the motion vector information. No other changes were made that could alter the decoder's behavior or results.

TABLE I. Video Names and Resolutions

| Video Name | Resolution |
|---|---|
| BQSquare | 416x240 |
| BlowingBubbles | |
| Mobisode | |
| BasketballPass | |
| Horses | 832x480 |
| Keiba | |
| BasketballDrill | |
| PartyScene | |
| FourPeople | 1280x720 |
| KristenAndSara | |
| Parkrun | |
| Stockholm | |
| BasketballDrive | 1920x1080 |
| Kimono | |
| Parkscene | |
| Tennis | |

The test signals were grouped into 4 experimental sets that contain test signals of the same spatial resolution in order to be possible the relative comparison of their content dynamics

based on the proposed grid and metrics. Thus, based on this grouping, the following sets have been created:

### *A. Experimental Set of Signals with 416x240 resolution*

The first differentiation experiment is among four test signals with 416x240 resolution, namely *BasketballPass*, *BQSquare*, *BlowingBubbles* and *Mobisode*. The experimental resutls of the proposed method are depicted on Figure 5.

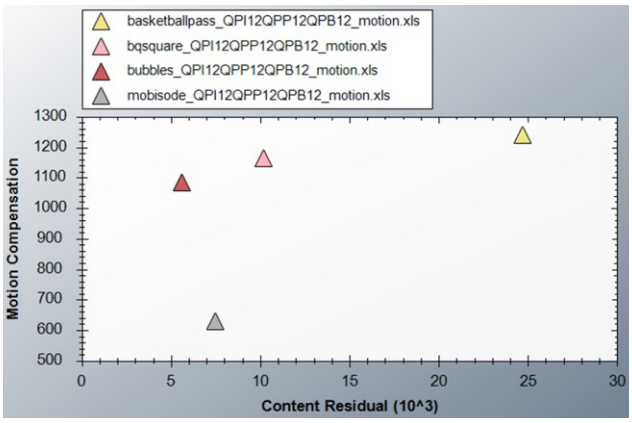


Figure. 5. Graphic Representation of the 416x240 video set.

Based on the results of Figure 5, it is evident that the video sequence *basketballPass* is the most active of the group in both the motion compensation and the content residual plane. This result is utterly justified by the video's content. The basketballPass sequence contains rich spatial and temporal motion activity as it comprises of several athletes moving constantly during a basketball practice. The overall motion activity is not only continuous during the duration of the video, but also very rich as it contains abrupt and fast movement from frame to frame, depicting the athletes' quick movements. Additionally, the camera alters its focus during the video, thus creating background changes which contribute to the overall mean motion vector number. All this quick and fast motion activity of the video elements are translated into a high valued mean motion vector length (content residual), but also into a large motion vector number (high valued motion compensation index), thus placing the *baketballPass* video signal on the top right corner relatively to the other three videos, which is mapped to both rich spatial and temporal motion activity.

Observing at the bottom left on the diagram, the *Mobisode* video sequence is located, which is placed as the video signal with the third largest motion vector length per frame, thus the second largest content residual index, but at the same time with the smallest motion compensation index. The *Mobisode* video sequence contains mainly homogeneous content, with a few scene changes and camera zooms, and without an overall high motion activity. So, its relatively mid valued content residual index is justified by the scene changes, in which a fair number of motion vectors shift or change completely in order to compensate for the successive image differences, but also from the camera zooms where large pixel areas are shifted. Due to the video signal's color, luma homogeneity and low motion dynamics, the encoded signal does not have a large number of motion vectors, as the color shifts in adjacent frames are minimal, thus the smallest valued motion compensation index.

The remaining two videos *BQSquare* and *BlowingBubbles* appear to be very similar spatio-temporally. Based on the results of the video differentiation method, they both have significantly richer content residual than the Mobisode sequence. Initially, the BQSquare sequence consists of a single shot recording on a square with several people doing various activities. The camera zooms out during the recording of the sequence which contributes to the growing number of motion vectors (motion compensation index) from frame to frame, as new objects enter and exit our perspective and contribute to the overall residual increase of the video signal. In the case of the *BlowingBubbles* sequence, the video sequence consists of two girls in a party blowing bubbles. The bubble artifacts create the majority of the motion activity in both axes, but their continuous shift and movement at many subsequent frames, generates a lot of motion compensation, thus the high valued motion vector number. However, their continuous slow movement throughout the whole video does not add up significantly to the mean motion vector length, thus the *BlowingBubbles* sequence has the lowest value of the group.

### *B. Experimental Set of Signals with 832x416 resolution*

The second evaluation set consists of four videos with 832x480 image resolution, namely the *BasketballDrill*, *Horses*, *Keiba*, and *PartyScene*. Figure 6 presents the visualized results provided by the proposed video differentiation process.

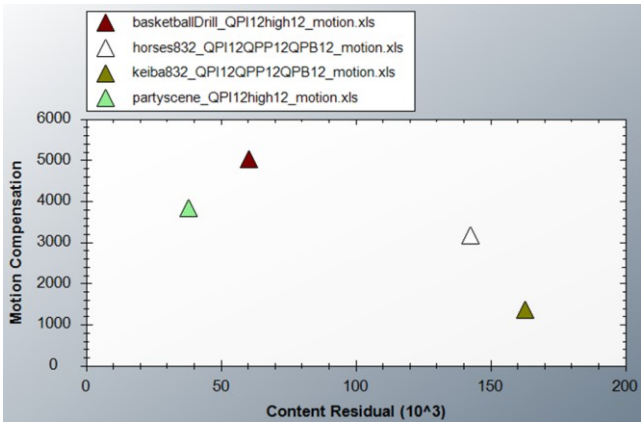


Figure. 6. Graphic Representation of the 832x480 video set.

The most distinct video differentiation position in this graph is the score of the *Keiba* video sequence. The *Keiba* signal is recorded during a horse race, where the camera is locked upon a single horse and follows it during the whole video. The motion dynamics in the video sequence are very high because of the fast background change, as the camera is locked-in, an abrupt and constant background change is observed. The horse's shift from frame to frame translates into high valued motion vectors, as the horse's pixels shift fast. The high valued mean motion vector length translates into a high valued content residual index. The main motion compensation activity is created by the background change as

the camera follows the horses, but the background has a repetitive pattern, which does not add up to the total motion vector number, and thus the overall motion compensation index.

Moving on the second rated most motion compensated video sequence under test, the *RaceHorses* video signal is noticed. The video *RaceHorses* appears to be the middle case among all the videos in the group under test, the video signal is a close up to a group of walking horses along with their riders. The rich color moving and the rapid horse movement generate rich spatiotemporal activity, along with the camera close-up. The camera zoom creates multiple divergent areas. The *RaceHorses* generates a similarly high valued mean motion vector length, as the Keiba video sequence. However, the video's homogeneous background and colors, create a smaller number of motion vectors than the other two videos, as they smoothly shift from frame to frame, thus not creating large pixel area movements, and a relatively low valued content residual index.

For the remaining two video signals *PartyScene* and *BasketballDrill*, it is observed that they have the highest motion compensation levels and the lowest content residual index. The *PartyScene* video signal consists of a few children playing and moving around in a room near a plant. The children's continuous slow paced movement results in low valued motion vector lengths and keeps the content residual index, at a low level. The camera zooms out during the whole video, thus the background changes (significant pixel area shifts) and this results in a high motion vector number and a high valued motion compensation index. In the last video, *BasketbalLDrill*, four basketball players perform a basketball exercise, by hitting one after the other against the basketball on the board. Similarly to the PartyScene video sequence BasketballDrill contains short macroblock shifts, due to the ball and player movement, which occur during whole video sequence. However, alike the PartyScene sequence the fast paced and continuous movement generates a large number of motion vectors. Which is why BasketballDrill has the highest valued motion compensation index.

### C. *Experimental Set of Signals with 1280x720 resolution*

The third evaluation set consists of four high definition videos of 1080x720 image resolution the *Stockholm*, *Parkrun*, *KristenAndSara*, and *FourPeople*. Figure 7 presents the visualized results for the video differentiation process applied to the aforementioned group of videos.

The visualized motion differentiation results for the third video set do not display significant motion compensation and content residual index differences among the current video sequences under-test, apart from the case of the *Stockholm* video signal which scores considerably higher on the content residual domain. The *Stockholm* video presents a panoramic overview of the city of Stockholm. The camera shifts in a fast pace on the horizontal axis and captures the ongoing activity of the city, mostly car traffic. The mean motion vector length (content residual domain) is relatively the highest in the group, due to the active motion activity. The shifting camera does not introduce new large moving objects and does not cause significant background change on the capturing frame, this generates a low motion vector number, thus holding the average motion vector number and the motion compensation index to its lowest value among the test set.

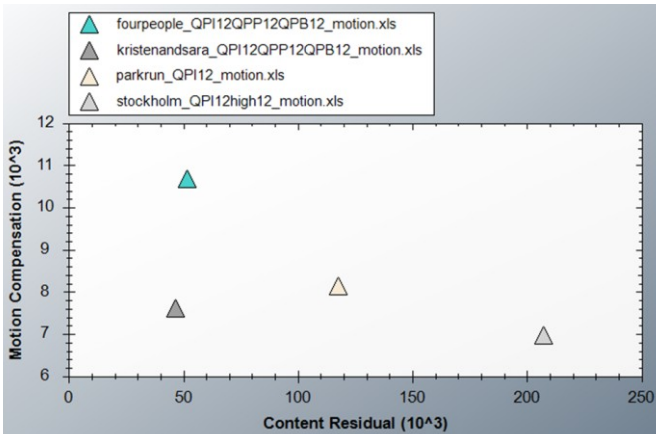


Figure. 7. Graphic Representation of the 1280x720 video group.

Heading at the center of the diagram, the next video, *Parkrun*, features a man running alongside a river in a park. The mean motion vector length is mainly generated by the movement of the man, but also from the fact that the river and the scene background record wide macroblock movements, due to the rich and highly detailed scenery. However, the detailed nature background, mainly because of the video's high definition provoke a considerable number of motion vectors, which adds up from frame to frame and finalizes in a relatively high average motion vector number per frame. Additionally, the continuously moving camera contributes a part to the increased number of motion vectors.

Finally, the remaining two videos *FourPeople* and *KristenAndSara*, both feature people talking, four and two people respectively. The two aforementioned videos, based on the differentiation process results, have the lowest content residual index. The *KristenAndSara* video content does not record overall a lot of active motion activity, as it only consists of two women having a chat, with mild body movement. Similarly, on the video *FourPeople* there are four people sitting around a conference table exchanging documents and having a discussion. In a similar manner, there is no fast, or abrupt motion recorded, which leads to the low motion vector length. Nonetheless, the average mean motion vector number is relatively high compared to the rest of the videos of the group under test. The reason why this occurs is that the camera is recording at a very close distance, so every movement is intensified, and this generates a high valued mean motion vector number. The persons' body movement is continuous throughout the video. The macroblock movement is not significantly active, as the camera is still (no background changes), thus the final low content residual index result for both videos.

### D. *Experimental Set of Signals with 1920x1080 resolution*

The fourth and final evaluation set consists of four true high definition videos of 1920x1080 image resolution the *ParkScene*, *BasketballDrive*, *Kimono*, and *Tennis*. Figure 8

presents the video differentiation process visualized results applied to the aforementioned group of video signals.

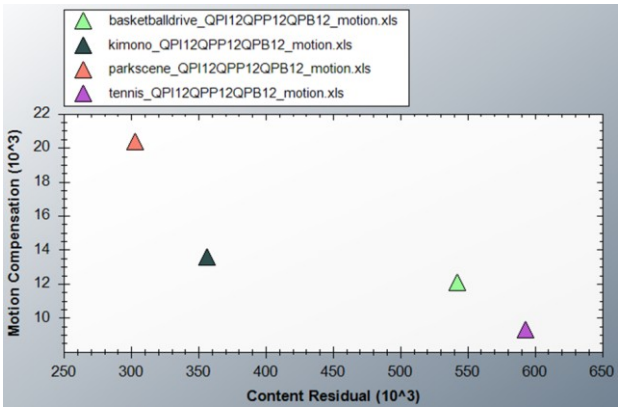


Figure. 8. Graphic Representation of the 1920x1080 video group.

In the last video set under test, it can be observed that the *Tennis* video sequence is by far the most motion active (content residual index is high) video sequence of the video group. The *Tennis* sequence features several tennis players preparing for a tennis match, and then records their activity during a play of the match. During the video the players continuously move across the terrain (generation of motion vectors) in a fast-paced manner, which translates into the high-valued content residual index of the video. However, the camera is recording from a distance, which means that the players' motion activity do not record a large number of motion vectors, thus the relatively lowest motion compensation index. The next most motion active video sequence is the *BasketballDrive* video signal, in this sequence we watch a basketball practice with similar non-stop movement during the duration of the video a static view prospect of the practice. The continuous fast movement creates high valued motion vector lengths, thus the high content residual index.

In the following sequence *Kimono*, we watch a lady pass through rich forest background and on the next scene she is approaching a pagoda building. The background in the first scene is very rich in detail and colors and changes significantly as the woman passes through. The background richness and the camera's zoom generate large number of motion vectors, as the tree-leaf pixels generate a lot of MB and pixel shifts. The relatively low content residual index is a result of the low motion dynamics of the video, as no significantly active movement is recorded. In the last video signal, the *ParkScene* namely, the video sequence features various people as they roam and meet each other in the park. The camera position is moving during the whole filming process generating a lot of motion vectors, due to background scenery change. The motion vectors are created by the bikers. It is justified that the *ParkScene* sequence has the highest valued motion compensation index as the background is rich and the general movement is continuous. But it has the lowest valued mean motion vector length as the bikers' and camera movement is slow paced and generates a low valued content residual index.

## VI. Conclusions

This paper proposes a novel video differentiation method based on the extraction of motion vector information. It is shown that the proposed technique, through the proposed metrics, achieves to capture and represent successfully the content dynamics of each video signal under test. Additionally, the proposed algorithm does not require highly complex calculations, but it can be easily embedded as an efficient content aware module into the existing video transmitting systems. More specifically, the proposed method can be efficiently exploited by the existing MPEG-DASH adaptation systems, where depending on the temporal, or the spatial activity of the transmitted video signal, the proposed method can provide an optimal decision for scaling up or down in the spatial or the temporal domain, depending on the spatiotemporal content dynamics of the test signal.

## Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 285621 (FP7 SEC SAVASA).

## References

[1] ISO/IEC 23009-1:2012 Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] I., Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet", IEEE Multimedia, Vol.18(4), 2011.

[3] N. Cranley and L. Murphy, "Incorporating User Perception in Adaptive Video Streaming Systems", in Digital Multimedia Perception and Design (Eds. G. Ghinea and S. Chen), published by Idea Group, Inc., May 2006. ISBN: 1-59140-860-1/1-59140-861-X

[4] "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC," in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050, 2003.

[5] Test Signals available online ftp://ftp.tnt.uni-hannover.de/testsequences