USING LATERAL RANKING FOR MOTION-BASED VIDEO SHOT RETRIEVAL AND DYNAMIC CONTENT CHARACTERISATION

Sergio Benini[1], Li-Qun Xu[2] and Riccardo Leonardi[1]

[1]University of Brescia, DEA, via Branze 38, 25123 Brescia, Italy
[2]BT Research, Adastral Park, Ipswich IP5 3RE, UK

{sergio.benini, riccardo.leonardi}@ing.unibs.it, li-qun.xu@bt.com

ABSTRACT

This paper investigates the issue of using a combination of motion descriptors, computed directly from MPEG motion vectors, for effective dynamic video content analysis and characterisation. On a shot by shot basis, the descriptors describe the general motion activities as well as spatial distribution of motions within a shot. On a frame by frame basis, they represent the continuous changes in pace and dynamics of the underlying video content. The former lends itself to an efficient motion-based shot retrieval scheme while being supported by a simple but effective lateral ranking fusion technique. It can be used for story segmentation and video summarisation too. And the latter is suitable for video skimming and fast video browsing. This system can be easily integrated with existing video retrieval and story segmentation system using only static visual features.

1. INTRODUCTION

Imagine you are working for an advertising agency, and a client asks you to create a motoring advert, promoting a new sports car model. You have already filmed scenes of the car racing along a rugged terrain and are keen to embed in between a sequence of shots depicting galloping horses (apparently not of the same colours or shapes etc) racing along similar path and with the same motion dynamics so as to enhance the compelling visual effects of the car to viewers. Given that you can access a large collection of video shots, you will now need to manually view and wade through the database to search for the required ones, which is most likely to be a tedious and labouring job. On the other hand, as a second scenario, your aim is to prepare an abstraction of the most exciting / dynamic moments for an action movie, given that you already have segmented scenes (obtained from visual-based techniques, e.g. [1]), the preferred choice is to select from each scene those shots that are most 'active', while the length used for each shot is proportional to its relative motion activity.

In this paper an attempt is made to address these and other similar requests, allowing for exploration of dynamic characteristics of video content to automate motion-based video shot indexing and retrieval as well as semantic content structuring. This complements other content analysis methods based on static visual appearance features including colour, shape and texture etc.

The proposed solution is based on the use of multiple motion descriptors to characterise the perceived motion activity as well as the unique spatial motion distribution in a given video segment. The descriptors used include the MPEG-7 motion activity descriptor [6], a motion intensity measure [15], the motion activity map [19], and dominant motion directions. All the descriptors are computed using compressed domain information extracted from P-frames. Once computed these descriptors provide effective shot-based metadata to support motion-based shot retrieval using query by example; a simple but effective lateral ranking method is introduced to fuse retrieval result from different descriptors. Experimental evaluations on News video, football content and a movie excerpt are carried out. The promising results obtained show that this scheme can be a valid support to a wide range of applications in the video content analysis domain, from similarity-based shot classification [5] to key-frame based video abstraction [4] and summarisation [3].

There is a great deal of interest in motion-based video content analysis research lately in both the compressed [7] and raw video domain [18][12]. The issues tackled include efficient motion characterisation methods/descriptors [2] [6], dominant /camera motion analysis [14][17], motion-based retrieval strategies [9][10][11], and motion-based video characterisation [8][13][16] etc. This paper contributes to this area of work by using multiple motion descriptors in a lateral ranking fashion that outperforms any single descriptor and by targeting at motion-based emerging application problems, e.g. video skimming.

The paper is organised as follows. Section 2 outlines the proposed system framework for motion-based dynamic content analysis; Section 3 then discusses the motion descriptors used to characterise the dynamic signature of a frame and/or shot. Section 4 introduces motion-based shot retrieval applications by lateral ranking, showing the results for news, sports video. Section 5 explains the applications for supporting content characterisation. The paper concludes in Section 6.

2. SYSTEM FRAMEWORK

The proposed system framework for motion-based dynamic content analysis is shown in Figure 1. It takes as input an MPEG-1/2 compressed video stream, and through partial decoding, exploits a number of low-cost descriptors computable directly from block motion vectors (MVs) and macroblock type attributes (MBT) extracted from each *P*-frame. These descriptors inherently describe the spatial-temporal changes of the motion properties of the video.

After suitable filtering to reduce typical noises due to block-based motion estimation, five motion descriptors are computed for each defined video segment. While some of these measures can be attached to either a frame or a shot, such as the statistical analysis performed on the magnitude of Motion Vectors, the Motion Intensity Descriptor [15] and the Dominant Motion Directions, others are shot-based only, including Motion Activity Map [19] and the MPEG-7 Motion Activity Descriptor [6].

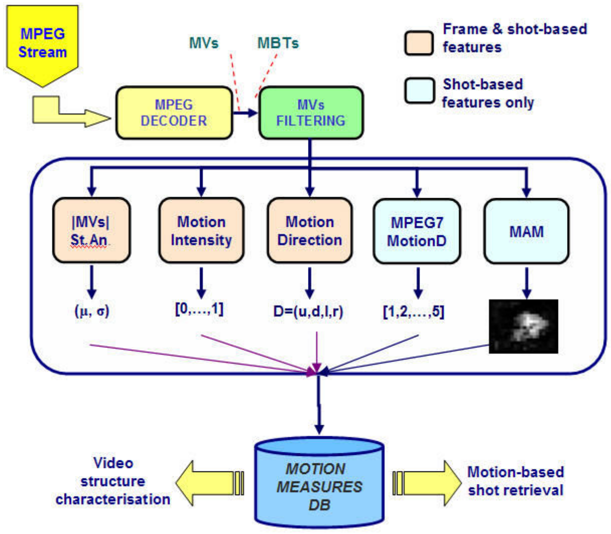


Figure 1: Schematic diagram of motion-based dynamic video content analysis system.

The shot-based motion descriptors can then be used for motion-based shot retrieval from a shot database (i.e. finding matching shots of similar motion patterns to that of a *query* shot) or searching for similar dynamic shots in a video sequence. The frame-based descriptors can be used to characterise the temporal structure of a video (e.g., to highlight the events in a football match or as a support to video summarisation). For the former an effective lateral ranking procedure is introduced to fuse the results from individual retrieval outcome.

3. MOTION FEATURES ANALYSIS

We now discuss in detail the motion features stated in Section 2 and the corresponding extraction methods adopted. As we know, for compression efficiency, MPEG uses a motion-compensated prediction scheme to exploit temporal redundancy inherent in an image sequence. In each GOP (group of picture) *I*-frames are used as references for the prediction. *P*-frames are coded using motion-compensated prediction from a previous *P* or *I*-frame (forward prediction) while *B*-frames are coded by using past and/or future pictures as references. This means that, in order to reduce the bit-rate, macroblocks (MBs) in *P* and *B*-frames are coded using their differences with corresponding reference MBs, and a motion vector carries the displacement of the current MB with respect to a reference MB.

The motion vectors decoded from a compressed video are normally coarse, noisy, and erratic; they may not be suited for tasks such as accurately segmenting moving objects, but can be very useful to characterise the general motion dynamics of a video sequence. Moreover, working in the compressed domain allows low-cost computation of motion measures, speeding up the processing of large amount of data for video indexing while offering fast query by example retrieval performance.

3.1. Stream Decoding

The desired output of the partial decoding of an MPEG video stream includes, for each P-frame, a sparse motion vector (MV) field (see, e.g., Figure 2) and a macroblock type (MBT) attribute attached to each decoded MB.

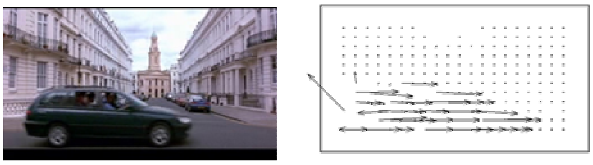


Figure 2: Illustration of a *P*-frame and the motion vector field extracted.

MBTs are needed since normally not all macroblocks in a *P*-frame carry motion information. For a macroblock using no motion compensation, i.e. a *No_MC* macroblock, it is further distinguished by an *inter/intra* classifier into two kinds of *No_MCs*: one is the *No_MC* intra-coded and the other is the *No_MC* inter-coded (as shown in Figure 3). The *inter/intra* classifier compares the prediction error with the input picture elements and, if the mean squared error of the prediction exceeds the mean squared picture element value, then the macroblock is intra-coded, otherwise, it is inter-coded. For our purposes it is important to note that *intra*-coded macroblocks, which do not have a coded motion vector, are assigned a zero motion vector.

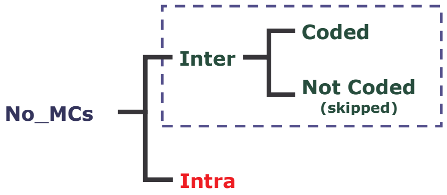


Figure 3: Distinction of two types of No_MC MBs in a P-frame. Zero motion is assigned to the intra-type.

3.2. Motion Vectors Filtering

The raw MV field extracted turns out to be very noisy and needs to be properly filtered prior to being used for motion descriptors analysis.

The chain of filtering process applied to raw motion vectors includes firstly removing MVs next to image borders which tend to be errant, then using a texture filter, followed by a median filter. The texture filter is needed since, in the case of low-textured uniform areas, the correlation methods used to estimate motion often produce spurious MVs. After having filtered the motion vectors on texture criteria, a median filtering is used to straighten up single spurious vectors such as those that could still be present close to the borders.

3.3. Statistical Analysis on Motion Vectors Magnitudes

Noting that the perceived visual activity is higher when the objects present in the scene move faster, or the magnitudes of the MVs of those macroblocks that make up the objects are significant, one simple measure of the global motion activity in a *P*-frame is provided by:

$$\mu = \frac{1}{\# MB} \sum_{i=1}^{\# MB} |MV|_{MB_i}$$

where μ is the mean of magnitudes of motion vectors belonging to *inter*-coded macroblocks only (Note that the *intra*-coded macroblocks have a zero motion vector).

Suppose we have a video shot that contains no moving objects, but it is captured by a camera in uniform motion. In this case the shot would be considered by most human viewers as a ‘non-active’ shot. This leads to the assumption that most of the perceived intensity is due to objects which do not move according to a uniform motion, thus the perceived motion activity is higher in the case of non-uniform motion. A good measure of the perceived motion intensity is the standard deviation σ:

$$\sigma = \left(\frac{1}{\# MB} \sum_{i=1}^{\# MB} \left(|MV|_{MB_i} - \mu \right)^2 \right)^{\frac{1}{2}}$$

The MPEG-7 motion activity descriptor discussed shortly is based on this standard deviation of MVs magnitudes.

The above two measures, computed for each *P*-frame, can be extended to characterise a meaningful video unit (e.g., a shot) as well by computing the mean and standard deviation of all the *P*-frames belonging to a unit.

3.4. MPEG-7 Motion Activity Descriptor

The intensity of motion activity is a subjective measure of the perceived intensity, or amount of motion activity in a video segment. For instance, while an ‘anchorman’ shot in a News program is perceived by most people as a ‘low intensity’ action, an ‘ice hockey’ game or a ‘car chasing’ shot would be viewed by most viewers as a ‘high intensity’ sequence.

The MPEG-7 motion activity descriptor described in [6] tries to capture the human perception of the ‘intensity of action’ or the ‘pace’ of a video segment. Note that it considers the overall intensity of motion activity in the scene, without distinguishing between the camera motion and the motion of the objects present in the scene.

The MPEG-7 motion activity descriptor uses quantised standard deviation of motion vectors to classify video segments into five classes, as shown in Table 1, ranging from ‘very low’ to ‘very high’ intensity.

Table 1: MPEG7 Motion Activity values

Range of MV st. dev. σ	MPEG-7 Motion Activity
$0 \leq \sigma < 3.9$	1
$3.9 \leq \sigma < 10.7$	2
$10.7 \leq \sigma < 17.1$	3
$17.1 \leq \sigma < 32$	4
$32 \leq \sigma$	5

3.5. Motion Intensity Descriptor

This feature is extracted according to [15] in its non-quantised version. From the definition of *inter No_MCs*, it is noted that when the video content changes slowly (‘low intensity’ action), or many macroblocks find a good match with their reference frames, the number of *inter No_MC* macroblocks is relatively large. On the other hand, when the video content changes quickly (‘high intensity’ action), the number of *inter No_MC* macroblocks is small, since many *MBs* cannot find a match in their reference frame, thus being directly *intra*-coded. So, the α-ratio of a *P*-frame can be defined as,

$$\alpha = \frac{\text{\# of inter NO\_MC MBs}}{\text{Total \# of MBs in the P - frame}}; \qquad 0 \leq \alpha \leq 1,$$

meaning that the higher the ratio is, the lower ‘intensity’ action is in the *P*-frame. A μ-law logarithmic compandor $G_\mu(\alpha)$ is given by:

$$G_{\mu}(\alpha) = Q \frac{\log(1+\mu\alpha/\mathrm{Q})}{\log(1+\mu)}; \qquad \alpha \le Q$$

It has been demonstrated in [15] that the obtained value is a good measure of the scene motion intensity and it conforms reasonably well to human perception.

Again, this measure, which is computed on a single *P*-frame, can be extended to characterise a meaningful video unit by averaging over all the *P*-frames within this unit. The mean value, together with the standard deviation, can then be used to characterise the unit's motion intensity; this measure is not depending on the video unit's size, therefore allowing also comparisons at multiple video levels.

3.6. Motion Activity Map (MAM)

At times when we are concerned with global motion changes instead of individual objects moving in the scene, we can view the motion of a video segment from the image plane along its temporal axis, as suggested in [19], by generating the so-called MAM (motion activity map).

The value of each pixel *(i,j)* in image $\mathrm{I}_{\mathrm{MAM}}$ is the numeric integral of the MVs magnitudes computed in its position and represents the measurement of the amount of motion during a period of time (from t_0 to t_f), i.e.:

$$I_{MAM}(i,j) = \frac{1}{t_f - t_0} \sum_{t=t_0}^{t_f} \left| MV(i,j,t) \right| .$$

In Figure 4, an example of the MAM is given, where the brightest regions correspond to the highest motion intensity zones. The utility of motion activity maps is twofold: on the one hand, it indicates if the activity is spread across many regions or restricted to one large region, showing a view of spatial distribution of motion activity. On the other hand, it expresses the variations of motion activity over the duration of the shot, displaying the temporal distribution of the motion activity.

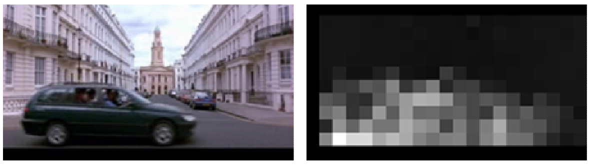


Figure 4: Motion Activity Map extracted from a shot.

3.7. Directions of Motion Activity

Besides the spatio-temporal motion properties described by a MAM, for a video shot that either contains several moving objects or is filmed by a moving camera, the approximate dominant motion directions can be very informative too [6]. Let MV_x and MV_y denote the two components of the motion vector MV of an MB, the total amount of motion along each of the four directions can be represented as a vector *D = (u, d, l, r)*:

$$u = \sum_{i=1}^{\#MB} (MV_y)_{MB_i} \quad if\ MV_y \le 0,$$
$$d = \sum_{i=1}^{\#MB} (MV_y)_{MB_i} \quad if\ MV_y > 0,$$
$$l = \sum_{i=1}^{\#MB} (MV_x)_{MB_i} \quad if\ MV_x > 0,$$
$$r = \sum_{i=1}^{\#MB} (MV_x)_{MB_i} \quad if\ MV_x \le 0.$$

A vector *D* is computed for each *P*-frame; it is then straightforward to extend this measure to characterise a shot by computing the mean and standard deviation of all *P*-frames contained.

4. MOTION-BASED SHOT RETRIEVAL

Given a query shot, in order to search for similar shots in a shot database that match its motion activity descriptors, we need to define a 'shot-to-shot' similarity distance measure for each of the five descriptors discussed before, or d_i, *i={1, 2,…,5}*:

- d_1: *L_1-norm* distance between standard deviations of motion vector magnitudes *|MVs|* (*σ*);
- d_2: *L_1-norm* distance between companded versions of *α*-ratio motion intensity descriptor (*$G_\mu(\alpha)$*);
- d_3: *L_1-norm* distance between vectors of dominant motion direction (*D(u, d, l, r)*);
- d_4: distance between two MAM images, defined as:
  $$d_4(I_{Sq}, I_{Sr}) = \sum_i \sum_j \left| I_q(i,j) - I_r(i,j) \right|$$
  where I_{Sq} and I_{Sr} are the MAM associated to the *query shot* s_q and the *reference shot* s_r, respectively.
- d_5: *L_1-norm* distance between *MPEG*-7 Motion Activity Descriptor.

4.1. Simple Lateral Ranking

Let *X={s_j}, j=1, 2,…, N,* be the shot database. Given a query shot s_q, the evaluation of the distances *{d_i}* between s_q and the shots in *X* produces one ranked list l_i for each considered descriptor, with the best matching shots (i.e. those at minimum distance d_i) in the very first positions.

Let the ranked list l_i with respect to distance d_i be:

$$l_i = \{(s_j)_{r_i}\}$$

where *j=1, 2,…N* is the shot index in *X*, and *r_i=1, 2,…, N* is the shot ranking position in the considered list l_i. To build a unique list *L* with the best matching shots we consider a simple lateral ranking procedure. This means that for each shot s_j in *X* we compute its total score *S(s_j)* by summing all rank positions relative to each list, i.e.:

$$S(s_j) = \sum_{i=1}^{5} r_i$$

The final list *L* will contain all shots *{s_j}* belonging to *X*, ranked by ascending order of their score *S(s_j)*.

This simple lateral ranking method turns out to be very effective since it allows combining the retrieval results from multiple features using different metrics, resulting in a unique ranked list. An example of shot-retrieval with the creation of the final ranking list *L* from separate lists l_i is shown in Figure 5.

4.2. News Videos

A software prototype for motion-based shot retrieval is realised based on preceding discussions. The first set of experiments is carried out using the video shots from a 48-minute Portuguese News programme.

The test set comprises a total of 476 shots ranging from low to high motion activities (manual annotation is used to verify the retrieval results). For each query shot, the best 30 similar shots are returned in the form of a ranking list, taking less than 1 sec on a modern PC.

Examples of query shots and the best matching ones retrieved are displayed in Figure 6, where the 'low' motion query shots are number 11, 346, 460 and the 'high' motion ones 435, 441, 165, 309. Some minor retrieval errors in the case of shots with significant motions are observed. For this application, only the distances from d_1 to d_4 are considered; the MPEG-7 Motion Activity Descriptor fails to produce a good rank due to its quantisation nature.

4.3. Football Videos

As an extension of the preceding experiments, the system is applied to football videos with shot type classification purposes. A 20-minute football match is used, consisting of 134 shots in total.

For evaluation purpose, all the video shots are manually annotated into three types, i.e. 'bird's eye view', 'medium' and 'close-up'. Given a query shot, the test aims to retrieve shots belonging to the same type in the very first positions of the ranked list.

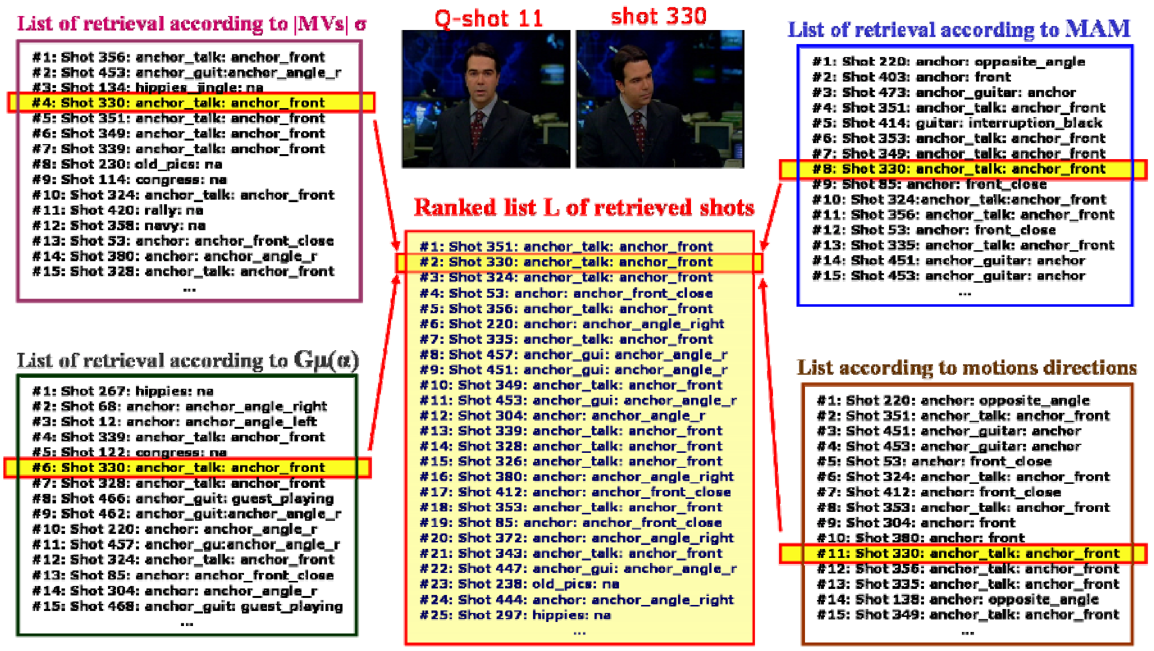
In general, we observed that the retrieval system performs quite well based on motion characteristics alone, though errors may occur on 'medium' type of shots, due to its wider range of motion variations, while good results are obtained on classifying 'close-up' shots. Figure 7 shows a retrieval example for each of the shot types.

5. SUPPORT FOR VIDEO CHARACTERISATION

In this section we explain how the motion-based shot retrieval framework can be a valid support for two different video characterisation tasks.

5.1. News Stories Segmentation

In a news programme, the 'news stories' (e.g., reports) are usually delimited by 'anchorperson' shots, as shown in Figure 8 (even though not all anchor-shots are always story boundaries!); moreover, all anchor-shots present similar motion patterns, i.e., they are filmed with a static camera, with little object motion and usually in a single broadcast filming technique, producing visually similar results even with different anchorpersons.



List of retrieval according to |MVs| σ

```
#1: Shot 356: anchor_talk: anchor_front
#2: Shot 453: anchor_guit:anchor_angle_r
#3: Shot 134: hippies_jingle: na
#4: Shot 330: anchor_talk: anchor_front
#5: Shot 351: anchor_talk: anchor_front
#6: Shot 349: anchor_talk: anchor_front
#7: Shot 339: anchor_talk: anchor_front
#8: Shot 230: old_pics: na
#9: Shot 114: congress: na
#10: Shot 324: anchor_talk: anchor_front
#11: Shot 420: rally: na
#12: Shot 358: navy: na
#13: Shot 53: anchor: anchor_front_close
#14: Shot 380: anchor: anchor_angle_r
#15: Shot 328: anchor_talk: anchor_front
...
```

List of retrieval according to Gμ(α)

```
#1: Shot 267: hippies: na
#2: Shot 68: anchor: anchor_angle_right
#3: Shot 12: anchor: anchor_angle_left
#4: Shot 339: anchor_talk: anchor_front
#5: Shot 122: congress: na
#6: Shot 330: anchor_talk: anchor_front
#7: Shot 328: anchor_talk: anchor_front
#8: Shot 466: anchor_guit: guest_playing
#9: Shot 462: anchor_guit:anchor_angle_r
#10: Shot 220: anchor: anchor_angle_r
#11: Shot 457: anchor_gu:anchor_angle_r
#12: Shot 324: anchor_talk: anchor_front
#13: Shot 85: anchor: anchor_front_close
#14: Shot 304: anchor: anchor_angle_r
#15: Shot 468: anchor_guit: guest_playing
...
```

Ranked list L of retrieved shots

```
#1: Shot 351: anchor_talk: anchor_front
#2: Shot 330: anchor_talk: anchor_front
#3: Shot 324: anchor_talk: anchor_front
#4: Shot 53: anchor: anchor_front_close
#5: Shot 356: anchor_talk: anchor_front
#6: Shot 220: anchor: anchor_angle_right
#7: Shot 335: anchor_talk: anchor_front
#8: Shot 457: anchor_gui: anchor_angle_r
#9: Shot 451: anchor_gui: anchor_angle_r
#10: Shot 349: anchor_talk: anchor_front
#11: Shot 453: anchor_gui: anchor_angle_r
#12: Shot 304: anchor: anchor_angle_r
#13: Shot 339: anchor_talk: anchor_front
#14: Shot 328: anchor_talk: anchor_front
#15: Shot 326: anchor_talk: anchor_front
#16: Shot 380: anchor: anchor_angle_right
#17: Shot 412: anchor: anchor_front_close
#18: Shot 353: anchor_talk: anchor_front
#19: Shot 85: anchor: anchor_front_close
#20: Shot 372: anchor: anchor_angle_right
#21: Shot 343: anchor_talk: anchor_front
#22: Shot 447: anchor_gui: anchor_angle_r
#23: Shot 238: old_pics: na
#24: Shot 444: anchor: anchor_angle_right
#25: Shot 297: hippies: na
...
```

List of retrieval according to MAM

```
#1: Shot 220: anchor: opposite_angle
#2: Shot 403: anchor: front
#3: Shot 473: anchor_guitar: anchor
#4: Shot 351: anchor_talk: anchor_front
#5: Shot 414: guitar: interruption_black
#6: Shot 353: anchor_talk: anchor_front
#7: Shot 349: anchor_talk: anchor_front
#8: Shot 330: anchor_talk: anchor_front
#9: Shot 85: anchor: front_close
#10: Shot 324:anchor_talk:anchor_front
#11: Shot 356: anchor_talk: anchor_front
#12: Shot 53: anchor: front_close
#13: Shot 335: anchor_talk: anchor_front
#14: Shot 451: anchor_guitar: anchor
#15: Shot 453: anchor_guitar: anchor
...
```

List according to motions directions

```
#1: Shot 220: anchor: opposite_angle
#2: Shot 351: anchor_talk: anchor_front
#3: Shot 451: anchor_guitar: anchor
#4: Shot 453: anchor_guitar: anchor
#5: Shot 53: anchor: front_close
#6: Shot 324: anchor_talk: anchor_front
#7: Shot 412: anchor: front_close
#8: Shot 353: anchor_talk: anchor_front
#9: Shot 304: anchor: front
#10: Shot 380: anchor: front
#11: Shot 330: anchor_talk: anchor_front
#12: Shot 356: anchor_talk: anchor_front
#13: Shot 335: anchor_talk: anchor_front
#14: Shot 138: anchor: opposite_angle
#15: Shot 349: anchor_talk: anchor_front
...
```

Figure 5: Simple lateral ranking. Given a query shot (shot 11), the best matching shots are presented in the ranked list *L*. Note that the first few retrieved shots in *L* correctly present the 'anchorman.' The high position in each list l_i corresponding to individual motion measures is also shown for shot 330.

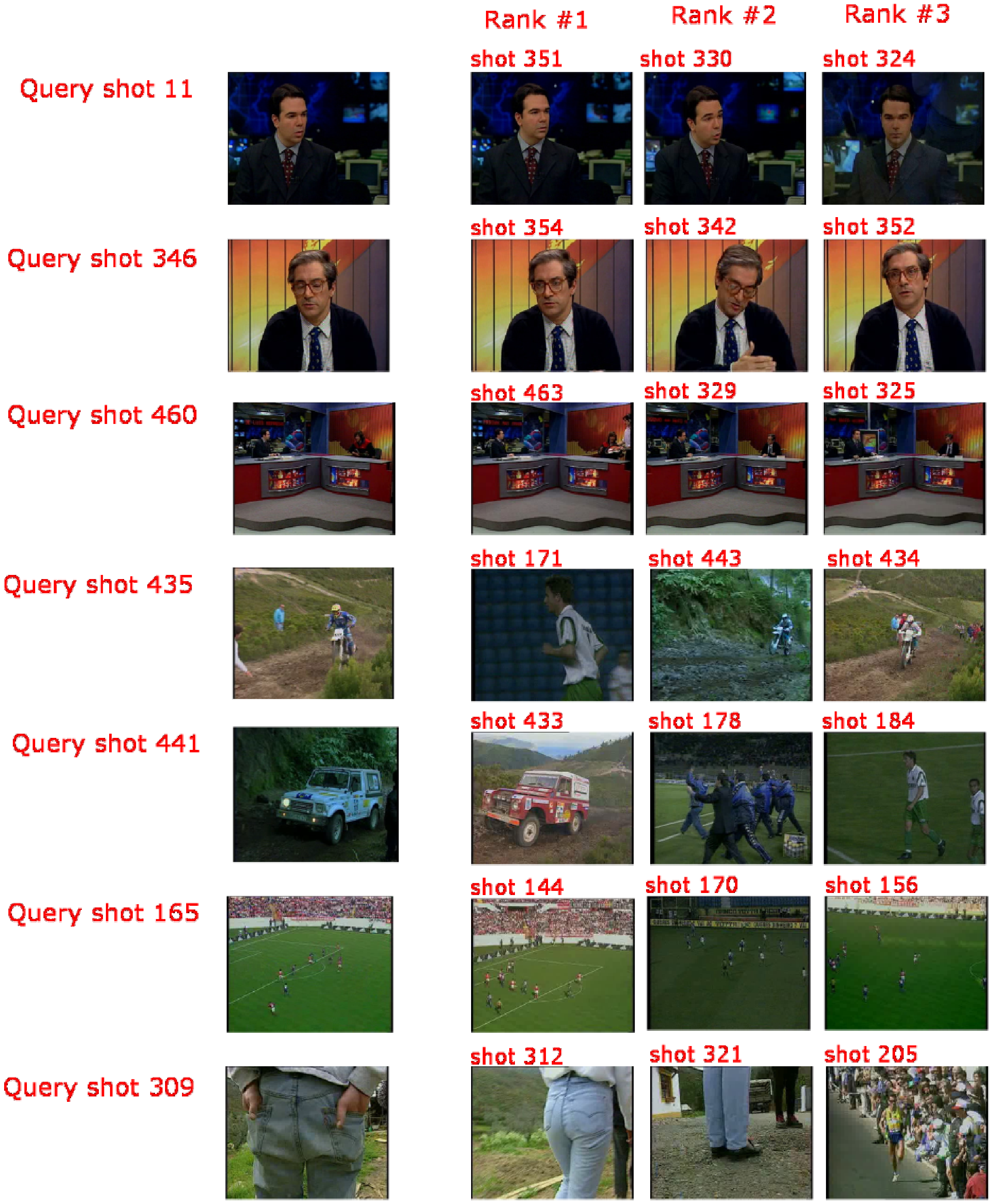


Figure 6: Examples of motion-based shot retrieval: 'low' activity scenes (query shot 11, 346, 460) - anchorperson, interviews; 'high' activity shots (query shot 435, 441, 165, 309) - motor biking, motor racing, football matches, close-up motions. The retrieved shots on the right exhibit similar motion characteristics to those of the query shots, respectively, even though they may show different visual contents.

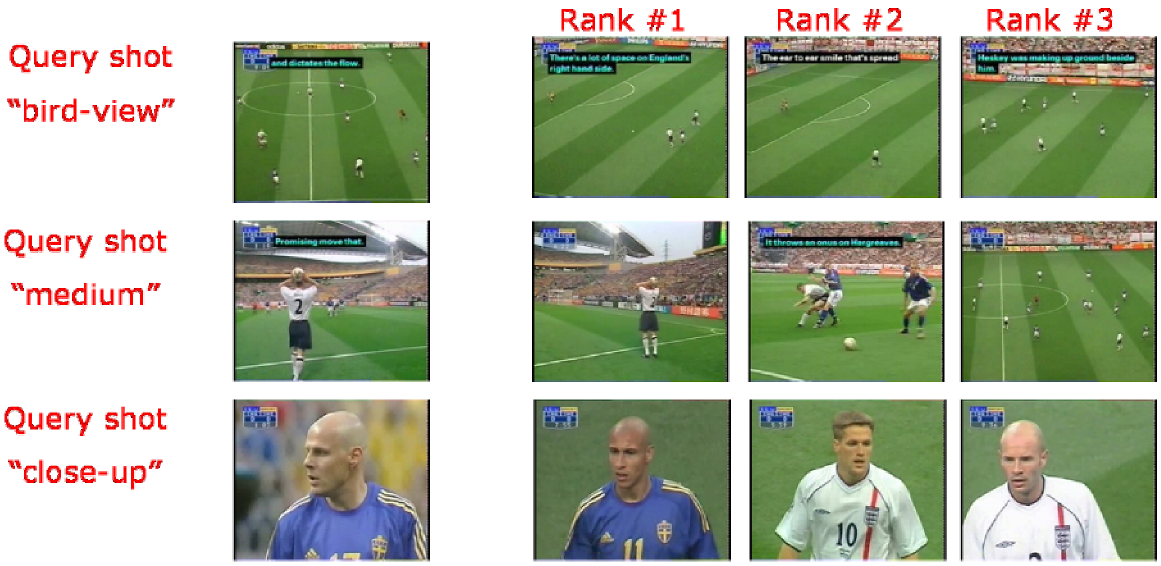


Figure 7: Motion patterns can be a support for sport shot classification into 'bird's eye view,' 'medium' and 'close-up'. Given a query shot, the system retrieves shots of the same type in the very first positions of the ranked list.

Efficient retrieval by our motion activity analysis framework allows identifying anchorperson shots for supporting segmentation of 'news stories'. Experiments on the Portuguese News video (where 37 among 476 shots are anchorperson shots) show that, given an anchorperson query shot, 28 anchor-shots are retrieved in the first 30 positions of the ranked list *L*.

Story 1
Story 2
Story 3

Figure 8: News stories delimited by anchor-shots.

5.2. Movie Summarisation

As stated in [3] the 'high' or 'low' intensity of action is in fact a measure of 'how much' the content of a video is changing. So motion activity can be interpreted as a measure of the 'entropy' of a video segment, and can be used for summarisation purposes.

As a result, possible applications of motion analysis range from highlight detection in sports video to key-frame selection for static video summaries [3] as well as building dynamic summaries (video skims), which we describe next.

As an extension of the work in [1] where shots have been grouped in semantically coherent *logical story units (LSUs)*, we can now investigate where inside an *LSU* the rapid video content changes occur, or identifying those high entropy shots. In an initial study we compute the frame-based measure of the standard deviation σ of Motion Vector Magnitude (duly filtered with a Kaiser-window of 40 seconds long) for a 10 minutes excerpt from the movie 'Notting Hill' (143 shots), which is plotted in Figure 9. We can see the close links, inside each *LSU*, between the high peaks in the σ-curve and the key-points in the content change of the story-plot.

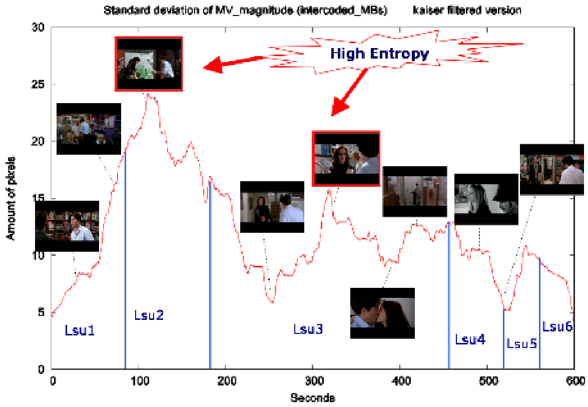


Figure 9: Plot of the smoothed σ-curve of the MV magnitude for a 10 min excerpt from the movie 'Notting Hill'.

As a video skimming application, given that a user can choose the length of the skim he or she would like to watch (subject to a minimal length). We propose that the time allocation policy for realising the skim should take into account the following criteria:

- the skim will have each *LSU* represented so as to include each semantically coherent story segment;
- each *LSU* will receive an allocation of the available time-length proportional to its duration against the whole movie;

- inside each *LSU*, only shots with higher entropy will enter the skim, in order to capture most 'informative' segments (i.e. the shots in which the visual content changes most rapidly).

This idea for effective video skimming is illustrated in Figure 10.

6. CONCLUSIONS

We have described an effective motion-based analysis framework for dynamic video content retrieval and characterisation. The system works efficiently in MPEG compressed domain on partially decoded motion information. Using a lateral ranking procedure, the dynamic shot retrieval task is achieved by fusing the results from five different shot-based motion descriptors. It has also been shown that the motion descriptors can be used to characterise continuously the temporal structure of a long video useful for story summarisation and video skimming. Future work is to integrate this with other video retrieval and story segmentation system that currently only use visual appearance features.

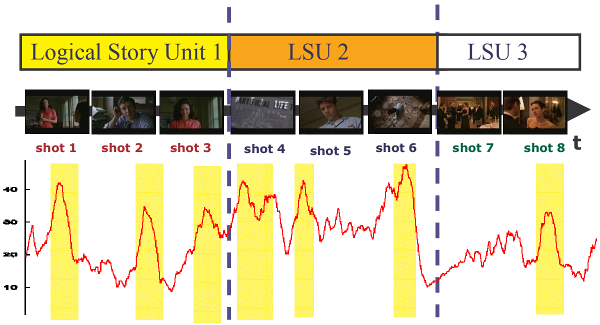


Figure 10: Video skimming: for each *LSU* time is assigned to 'high intensity' motion segments.

7. REFERENCES

[1] S. Benini, L.-Q. Xu and R. Leonardi, "Identifying video content consistency by Vector Quantization," *Proc. of WIAMIS'05*, Montreux, Switzerland, April 2005.

[2] E. Bruno, D. Pellerin, "Video structuring, indexing and retrieval based on global motion wavelet coefficients," *Proc. of ICPR'02*, Quebec City, Canada, August 2002.

[3] A. Divakaran, K. A. Peker, et al., "Video summarization using mpeg motion activity and audio descriptors," *MERL TR2003-034*, May 2003.

[4] A. Divakaran, R. Radhakrishnan, K. Peker, "Motion activity-based extraction of key-frames from video shots," *Proc. of IEEE ICIP'02*, Sept. 2002.

[5] B. Fablet, P. Bouthemy, P. Pérez, "Statistical motion-based video indexing and retrieval," *Conf. on Content-Based Multimedia Information Access*, (RIAO'2000), Paris, April 2000.

[6] S. Jeannin, A. Divarakan, "MPEG-7 visual motion descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, **11**(6), June 2001.

[7] N.W. Kim, T.Y. Kim, J.S. Choi, "Motion analysis using the normalization of motion vectors on MPEG compressed domain," *ITC-CSCC2002*, **3**, pp. 1408-11, July 2002.

[8] Y. Li, L.-Q. Xu, G. Morrison, C. Nightingale, J. Morphett, "Robust panorama from MPEG video," *Proc. of IEEE ICME'03*, Baltimore, July 2003.

[9] H. Lu and Y.P. Tan, "Sport video analysis and structuring," *Proc. of IEEE MMSP'01*, Cannes, France, Oct. 2001.

[10] Y.-F. Ma, H.-J. Zhang, "Motion pattern based video classification and retrieval," *Journal on Applied Signal Processing*, No. 2, Feb. 2003.

[11] V. Mezaris, I. Kompatsiaris, N.V. Boulgouris, and M.G. Strintzis, "Real-time compressed domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *IEEE Trans. on C.S.V.T,* **14**(5), May 2004.

[12] C.W. Ngo, T.C. Pong, H.J. Zhang, "On clustering and retrieval of video shots through temporal slices analysis," *IEEE Trans. on Multimedia*, **4**(4), Dec. 2002.

[13] K.A. Peker and A. Divakaran, "Framework for measurement of the intensity of motion activity of video segments," *J. of Visual Communications & Image Representation*, **14**(4), Dec 2003.

[14] M. Pilu, "On using raw MPEG motion vectors to determine global camera motion," *Technical report*, HP Laboratories, Bristol, August 1997.

[15] X. Sun, A. Divakaran, B.S. Manjunath, "A motion activity descriptor and its extraction in compressed domain," *Proc. of IEEE Pacific-Rim Conference on Multimedia* (PCM'01), Beijing, Oct. 2001.

[16] X. Sun, B.S. Manjunath, and A. Divakaran, "Representation of motion activity in hierarchical levels for video indexing and filtering," *Proc. of IEEE ICIP'02*, Rochester, NY, Sept. 2002.

[17] R. Wang, T. Huang, "Fast camera motion analysis in MPEG domain," *Proc. of IEEE ICIP'99*, Kobe, Japan, 1999.

[18] L.-Q. Xu, J. Zhu, and F. Stentiford, "Video summarization and semantic editing tools," *Proc. of SPIE Conference on Storage and Retrieval for Media Databases*, Vol. 4315, pp. 242-252, January 2001.

[19] W. Zeng, W. Gao, and D. Zhao, "Video indexing by motion activity maps," *Proc. of IEEE ICIP'02*, Rochester, NY, Sept. 2002.