

# Low-level Motion Activity Features for Semantic Characterization of Video

*Kadir A. Peker, A. Aydin Alatan, Ali N. Akansu*  
New Jersey Center for Multimedia Research  
New Jersey Institute of Technology  
E-mail: [kap3510@oak.njit.edu](mailto:kap3510@oak.njit.edu)

## Abstract

Efficient methods of content characterization for the browsing, retrieval or filtering of vast amount of digital video content has become a necessity. Still, there is a gap between the computationally available measures of content characteristics and the semantic interpretations of these characteristics. We want to establish connections between motion activity characteristics of video segments and the semantic characterization of them. For this purpose, two simple descriptors for motion activity of a video content is used to infer high-level semantic features of video in certain contexts. One of these descriptors, monotonous activity, is defined as the average block-based motion vector magnitude. The second descriptor, non-monotonous activity, is an approximation to the average temporal derivative of motion vectors. Simulation results for browsing and retrieval applications show that by using the two measures together, object motions that occur close to the camera can be distinguished from distant ones. Also by using the two descriptors together, we are able to differentiate between a high activity due to camera motion and a high activity due to dancing people. Hence, these simple descriptors, especially when used to complete each other, promise to provide important clues about semantics of a video.

## 1. Introduction

Digitization of audio-video content brings new opportunities for efficient utilization of this material in new and improved ways. The possibility of access to vast amounts of video content made possible by developments in communications and digital technology, such as high-speed digital networks, digital broadcasting, digital libraries and personal digital recorders, commend for efficient ways of organizing multimedia material, and thus the ability to retrieve or browse through desired segments of content. This requirement resulted in a large number of work in the area of content-based analysis of audio-visual material [1].

A first step in automatic or semi-automatic analysis of video is the segmentation of video into temporal units, such as shots, based on certain given criteria [2]. For example, a news program can be segmented into anchorperson shots, story footage, sports section, etc. as in [3]. Examples to more low-level temporal segmentations are indoors – outdoors shots, people

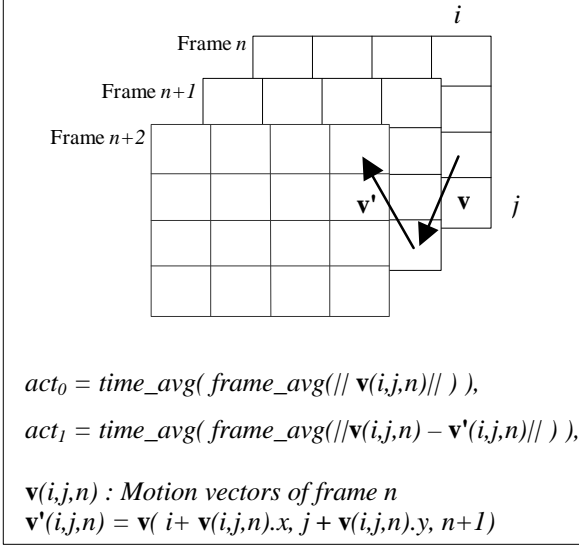
present – no people present, or close-up – wide-angle shots.

The problem to be solved in this case is the establishment of connections between computationally and practically available measures, and the semantically meaningful features of content[3].

Our approach is that there are connections between certain low level, computationally simple features and the high level, semantically meaningful features. Furthermore, these connections can be either intuitive and common sense (e.g. in a news program, anchorperson segments are usually characterized by very low activity compared to the other segments), or they can also be more subtle and indirect (e.g. close-up shots are characterized by unsteady motions, i.e. sharp changes in direction and magnitudes of motions in the scene, whereas wide angle shots mostly have steady, smooth global motions). We want to find these connections between certain high level characterizations of video and the corresponding low level measures that signal them. In particular, we are interested in analyzing various types of video segments in terms of their motion activity characteristics, and in finding a set of low-level motion activity measures that can reflect these characteristics.

Establishing such connections between low-level, computationally efficient measures and high-level characterizations of video is invaluable in real world applications where data volume is large and speed is critical. Especially in the area of content analysis, applications are the main driving forces and an important part of the definition of the problem itself. A limitation of the above approach, on the other hand, is the crudeness of content characterization. Thus, this approach is most appropriate for browsing or summarization tool, or as a preprocessing step before further analysis of the content.

In this paper, we demonstrate semantic inferences that can be made using two low-level motion activity features, or "descriptors" in MPEG-7 terminology. We define motion activity as the gross, overall motion content in a given video segment, such as when we say high or low activity, spatially coherent or scattered activity, etc. Both the descriptors that we use are implemented using block motion vectors from MPEG 1/2 coded streams. This allows for very fast and efficient implementations of the proposed methods. The first descriptor is the average of the magnitudes of block motion vectors in a given frame. The second is computed using the changes in block motion vectors



**Figure 1.** Computation of monotonous and non-monotonous activity descriptors from motion vectors.

from one frame to the next, and is a measure of "unsteady" or "non-monotonous" component of the motion activity. These measures reflect the intensity of activity in video segments, each emphasizing different types of activities, and their difference enables us to better understand the type of activity that is measured.

## 2. Motion Activity Descriptors

We describe two motion activity descriptors: Monotonous (steady) motion activity descriptor  $act_0$  and the non-monotonous (unsteady) motion activity descriptor  $act_1$ . The two descriptors are defined as follows:

$$act_0(n) = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^M \| (x(i, j, n), y(i, j, n)) \|$$

$$act_1(n) = \frac{1}{M \times N} \sum_{i=1}^N \sum_{j=1}^M \| (a(i, j, n), b(i, j, n)) \|$$

$$a(i, j, n) = x(i + dx, j + dy, n + 1) - x(i, j, n)$$

$$b(i, j, n) = y(i + dx, j + dy, n + 1) - y(i, j, n)$$

$$dx = \lfloor x(i, j, n) \rfloor$$

$$dy = \lfloor y(i, j, n) \rfloor$$

where,

$x(i, j, n)$  and  $y(i, j, n)$  are x and y components of the  $(i, j)^{th}$  block motion vector of frame n,

$M \times N$  : Size of the frame in macro - blocks,

$n$  : Frame number.

$act_0$  is a general measure of motion activity in the frame. It is computed as the average of the magnitudes of the motion vectors in the frame.

$act_0$  is sensitive to global motion such as camera pan and to objects moving very close to camera since these types of motions result in a high average of motion vector magnitudes.

$act_1$  has a compensating effect to the above sensitivity of  $act_0$ . It filters out the component of motion activity that does not change from frame to frame. Hence, a significant portion of steady global motions such as a camera pan or zoom, or a smooth translational motion of an object close to the camera is eliminated.  $act_1$ , in essence, is the average acceleration – change in both the magnitude and the direction of motion – of image points (or macro-blocks in our implementation) in a frame.

In contrast to  $act_0$ ,  $act_1$  is more sensitive to unsteady motions such as a hand held camera, or the fickle motions of a non-rigid object in close-up.

Frame based descriptors  $act_0(n)$  and  $act_1(n)$  can be extended to segment based descriptors by using a function like mean, median, max or min over the frames in the segment. We used the mean of frame activities in our implementations.

## 3. Results from Application Examples

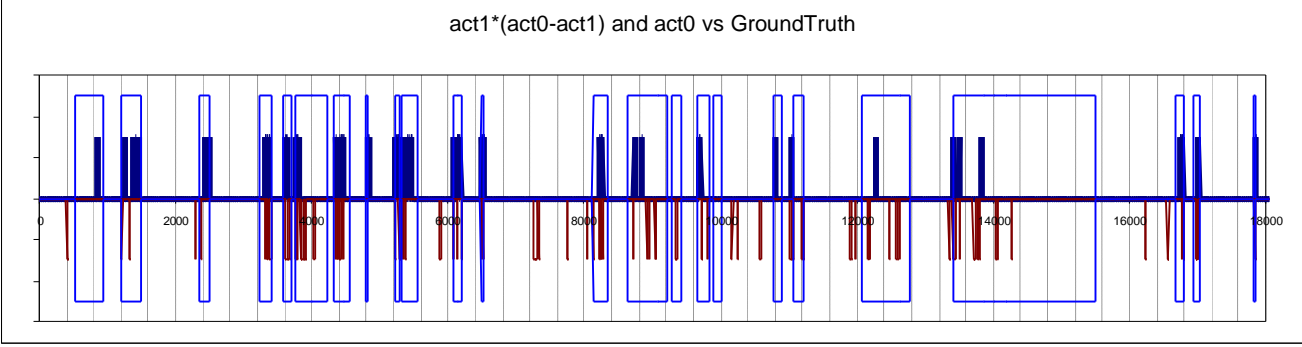
We used the two descriptors in two different application contexts: a) Browsing through a sports video (basketball and soccer from MPEG-7 data set), and b) Retrieval from a database of shots compiled from various programs such as news, sports, entertainment, education, etc. In the sports video example, we observed that the difference  $act_1(n) - act_0(n)$  is highest for close-up shots where the irregular motion of players in view is dominant over the regular global motion. In the retrieval example, we contrast a retrieval of shots with highest  $act_1(n) - act_0(n)$  difference to a retrieval based on high  $act_0(n)$  only. When we use  $act_0$  only, we get high activity shots resulting from camera motion or an object moving too close to the camera. But when we retrieve by using the two descriptors together, we are able to get the shots with dancing people, i.e. with high articulated, irregular motion and relatively less uniform global motion.

### 3.1 Detecting Close-ups in Sports Video

The two activity descriptors are computed for the P-frames of the basketball video from the MPEG-7 data set (10 minutes, 18000 frames, 4800 P-frames). A ground truth data is prepared manually, segmenting the video into wide angle and close-up shots. There were 59 segments, 30 of them being close-ups. Two measures are compared:

$m1 = act_0(n)$  and,

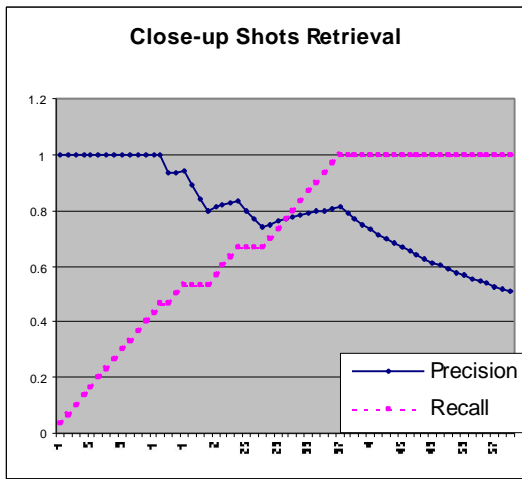
$m2 = act_1(n) \times (act_1(n) - act_0(n))$ .



**Figure 2.** Frame-based detection of close-ups using  $m1$  and  $m2$ . Bounding boxes are close-up segments. Positive impulses are where  $m2$  suggest a close-up, whereas negative impulses are where  $m1$  suggests a close-up. Note that  $m2$  impulses are always inside a close-up segment.

$m2$  is the difference mentioned above, scaled by the activity itself to emphasize the high activity segments. We expect that, if non-monotonous activity  $act_1(n)$  is significantly higher than  $act_0(n)$  in a frame, then with a high probability, the frame is a close-up on a highly active object. Also, we expect  $m1 (act_0(n))$ , as well, to be high for close-up frames because of the scaling up of motion vectors due to zoom.

The analysis is done first on a frame-based detection of close-ups. In Figure 2, manually extracted close-up segments are shown as bounding boxes. A threshold for  $m1$  and another threshold for  $m2$  is empirically selected such that 250 P-frames with highest  $m1$  and  $m2$  values are over the thresholds. Positive impulses are where  $m1$  is above the threshold, and the negative impulses are where  $m2$  is over the threshold. It is observed that the top 250 P-frames with respect to  $m2$  are consistently from close-up segments, whereas  $m1$  does not always guarantee that property. Thus,  $m2$ , in this example, functions as an indicator for close-up segments.



**Figure 3.** Precision-recall vs. the number of shots retrieved. Precision is 1 when we retrieve up to 16 shots.

In the second part of the analysis, segment based measures  $sm1$  and  $sm2$  are computed for each manually extracted segment of the video. The segment measures are based on frame-based measures  $m1$  and  $m2$ . In the previous frame-based experiment, we observed that when  $m2$  is over a threshold, than the segment is a close-up. We also know that  $m1$  is relatively higher for close-ups, since motion vectors are larger when the action is close to the camera. From these heuristics, we use first, the average of  $m1$  over the frames in the segment, and second, the number of frames in the segment for which  $m2$  is over the experimentally determined threshold. (As noted earlier, we choose the threshold such that 250 P-frames out of 4800 – about 5% – are above the threshold.)

$$sm1 = \frac{1}{N} \sum_{n=1}^N m1(n)$$

$$sm2 = \sum_{n=1}^N \partial 2(n)$$

$$\partial 2(n) = \begin{cases} 1 & \text{when } m2(n) > q \\ 0 & \text{otherwise} \end{cases}$$

$m1(n)$ :  $m1$  computed for frame  $n$

$m2(n)$ :  $m2$  computed for frame  $n$

$q$ : Empirical threshold to detect close - ups

$N$ : Number of frames in segment

We find the close-up segments by sorting the segments with respect to  $sm1$  and  $sm2$  in each experiment and choosing the top  $k$ . Figure 3 shows precision/recall computed for retrieved number of shots ( $k$ ) varying from 1 to 59, for retrieval of close-up shots using  $sm2$ . For  $k=16$ , all the retrieved segments are correctly close-up segments.

The retrieval using  $sm1$  (average of  $act0$  over the segment) gives a comparable performance overall but it is again misled by camera motion. The first retrieved segment, for example, is a fast pan segment. Hence, we find  $sm2$  to be a more reliable detector for close-ups.

### 3.2 Retrieval of High Activity Shots

In the second application, a database of 600 shots that are extracted from MPEG-7 test set is used. The database includes shots from a diverse set of programs such as news, sports, entertainment, education etc. 5 highest activity shots are retrieved using  $act_0$ ,  $act_1$ , and  $(act_0 - act_1)$ . The results are shown in Figures 4, 5 and 6.  $act_0$  and  $act_1$  retrieve shots that contain fast camera motions or an object that passes too close to the camera, which are not commonly considered high activity. However, by retrieving the shots that have high *non-monotonous* activity rather than *monotonous* activities such as camera pan or close range translational motion, we are able to get 5 shots of dancing people from the diverse set of 600 shots.

### 4. Conclusion

We described two descriptors for the characterization of the motion activity in a video segment. Simulations show that these two descriptors can be used to infer whether the activity content is dominantly a monotonous, steady motion or an unsteady, inconstant motion. Application examples demonstrate that this kind of a characterization of the activity content can

be used to detect close-up segments in a sports video, or in fine tuning of an activity based query from a database of video shots.

### 5. References

- [1] P. Aigrain, H. Zhang, D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-the-Art Review," *Multimedia Tools and Applications*, 3, pp. 179-202 (1996)
- [2] B.-L. Yeo and B. Liu, "Unified Approach to Temporal Segmentation of Motion JPEG and MPEG Video," Proc. International Conf. on Multimedia Computing and Systems, pp. 2-13, 1995.
- [3] B. Gunsel, A. M. Ferman, and A. Murat Tekalp, "Temporal Video Segmentation Using Unsupervised Clustering and Semantic Object Tracking," in J. of Electronic Imaging, vol. 7, no. 3, pp. 592-604, July 1998.
- [4] N. Vasconcelos, A. Lippman, "Towards Semantically Meaningful Feature Spaces for the Characterization of Video Content," *Proc. ICIP*, 1997.



Figure 4. 5 highest activity shots out of 600 when activity is measured by  $act_0$ .



Figure 5. 5 highest activity shots out of 600 when activity is measured by  $act_1$ .



Figure 6. 5 shots out of 600 that have highest  $(act_1 - act_0)$  difference. Note that,  $act_0$ , also  $act_1$  to some degree, is misled by camera motion. We are able to filter out high camera motion segments by using the  $(act_1 - act_0)$  difference.