Motion based video classification for SPRITE generation

*Abhidnya A. Deshpande and Ramazan S. Aygün**

Computer Science Department
University of Alabama in Huntsville
Huntsville, Al USA
{adeshpan, raygun}@cs.uah.edu

Abstract - In this paper we address the problem of video classification for sprite generation based on various features along with the global and local motion present in the video. Our feature set consists of features such as global (or camera) motion, cumulative global motion, local motion (motion of objects in the video), duration of the video, number of objects in motion, number of macro-blocks in motion and presence of objects at the borders of the image. These features are analyzed together to classify the video into one of the six pre-defined classes. The main focus of our approach is to analyze the number of frames that are processed in order to extract the feature set from the video. We perform experiments on a variety of videos by varying the number of frames being processed and analyze the outcome while calculating the accuracy of our approach.

Index Terms — Video processing, sprite generation, video classification

1. INTRODUCTION

Although sprite coding was proposed for MPEG-4 Main profile [9], to the best of our knowledge, there is no commercial video encoder that supports MPEG-4 Main profile. We believe that one of the reasons for this is the domain of videos where sprites can be generated. In other words, not all videos are suitable for sprite generation. Although the new H.264 standard [10, 14] does not support sprite coding, we believe that future standards that are based on H.264 can benefit from sprite coding. In [15], it is shown that better compression ratios are achieved with sprite coding using H.264 compression.

We believe that videos should also be classified for sprite generation. In the past, videos were classified for different purposes such as genre classification [5], sports video classification [1], news video classification [7], rule-based classification [3], motion and contour based classification [12] and so on. Some of the features that are used for video classification include texture, shape, audio, length of the video clip in frames, the number of shots, average shot length in frames, color histogram etc. [3], [5], [6]. The most common features that are used in classification are motion and color features. For example, Xavier et al [1] use Hidden Markov Model for sports video classification using motion and color features. In [13], the authors propose a technique of extracting the motion features along with the color features from the compressed video domain. These features provided a good characterization of video in both spatial and temporal directions. The classification was based on the maximum likelihood principal of classification that uses the Hidden Markov Model. Kuhne, Tichter and Beier [12] proposed segmenting video objects and then classified mainly based on motion signals. To extract the motion information from the video, the 3D structure tensor is utilized as a source of integrating information from a number of consecutive video frames. Furthermore, an active contour model was developed for estimating the motion in the video, and then it was matched with the processed views of the prototypical objects stored in the database.

In our earlier work [12], we have presented a method on how to classify videos for sprite generation. In this paper, we introduce new features such as motion pattern and provide a better classification of videos. We analyze the effect of the number of frames being processed. Here, we focus on motion features with more detail than other previous approaches. The global motion is considered from frame to frame as well as cumulative for the video shot. We also investigate global motion patterns in this paper. We extract 6 prominent features from the video namely: presence of duration, global motion, cumulative global motion, local motion, number of objects in motion, number of macro blocks in motion and presence of objects at the borders of a frame. The videos are then classified based on global motion, local motion, and presence of objects and presence of pattern in global motion if any.

The organization of our paper is as follows. The following section describes the features that we used for classification. Section 3 describes our classification approach. Our data set and experiments are provided in Section 4. The last section concludes our paper.

* This research is funded by NSF IIS-0812307.

2. FEATURE EXTRACTION

Our feature set for classifying videos for sprite coding is composed of 5 features: *{G, C_g, N, P, M}*. These correspond to global motion *(G)*, cumulative global motion *(C_g)*, number of objects in motion *(N)*, presence of objects at the borders of a video frame *(P)* and the number of macroblocks in motion (*M*). We briefly explain these features.

2.1 Global Motion and Cumulative Global Motion

2.1.1 Global motion parameters (G)

The global motion estimation is the first step of sprite generation. The global motion might be equivalent to the camera motion if the moving objects are not large. The global motion estimation corresponds to estimation of motion parameters. The perspective motion has 8 parameters and the new coordinates of a pixel at (x, y) is computed as:

$$x' = \frac{m_0 x + m_1 y + m_2}{m_6 x + m_7 y + 1} \qquad y' = \frac{m_3 x + m_4 y + m_5}{m_6 x + m_7 y + 1} \qquad (1)$$

where (x', y') is the position of the same pixel in the reference frame (i.e., next frame). For affine motion, m_6=0 and m_7=0. For translational-zoom-rotation, m_4=-m_3, m_5=m_2, m_6 = 0, and m_7 = 0. For translational motion, m_2=1, m_5 =1, m_4=0, m_5 =0, m_6 =0 and m_7 =0.

2.1.2 Cumulative maximum global motion (C_g)

Frame-to-frame global motion analysis may not provide a good picture of actual global motion of a video shot. The magnitude of frame-to-frame global motion parameters for a panning camera might be similar to the global motion parameters of an earthquake video. We need to identify the cumulative global motion for a video shot. However, this may not be equivalent to the global motion between the first frame and the last frame of a video shot.

Consider Fig. 1 for cumulative global motion. The pixels that are marked correspond to the same locations with respect to the size of a frame. The distance between them on the sprite indicates the cumulative global motion between frame 0 and frame 298 with respect to the selected pixel. Our goal is to find the cumulative *maximum* global motion between any pair of frames in the video shot.

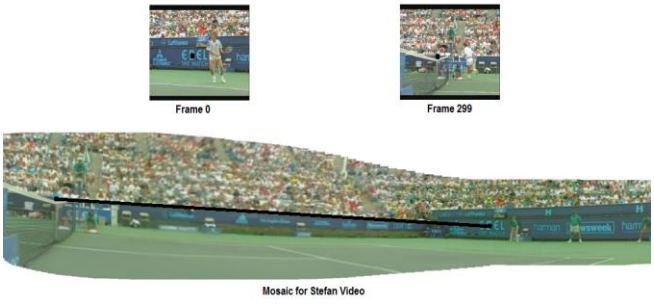

Figure 1: Cumulative Global Motion for a selected point

We need to choose a pixel location that is sensitive to any type of motion including affine, translation, and perspective and calculate the cumulative global motion with respect to this point. Figure 1 shows the first and the last frame for the Stefan video. It also highlights the chosen pixel position in each of these frames. Along with this it shows the generated mosaic for the Stefan video and highlights the maximal displacement of the chosen pixel position throughout the video.

In our approach the calculation of the cumulative global motion is implemented in 2 steps. Firstly we calculate the position of the selected pixel in all frames that are to be processed. Secondly, we calculated the distance between all pairs of these estimated pixel positions for the selected pixel and then choose the maximum value among them as the maximum cumulative global motion. The first step includes the matrix multiplication process. Following are the equations used to calculate the co-ordinates of the selected pixel in all frames:

$$\begin{bmatrix} m_0 & m_1 & m_2 \\ m_3 & m_4 & m_5 \\ 0 & 0 & 1 \end{bmatrix} X \begin{bmatrix} (w/4) \\ (h/4) \\ 1 \end{bmatrix} = \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \qquad (2)$$

where m_0, m_1, m_2, m_3, m_4, m_5 and m_6 are the transformation parameters, (w/4, h/4) is the chosen pixel position and (x_i, y_i) refers to the new position of the selected pixel in each frame. This chosen pixel provides a good measure of the cumulative global motion that is sensitive to rotation, translation and zoom. We used the Euclidian distance to measure the distance between pixels:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (3)$$

where (x_1, y_1) refer to the co-ordinates of the point in the current frame and (x_2, y_2) refer to the co-ordinates of the same point in the next frame. This distance provides the displacement of the chosen pixel position with respect to the first frame. The maximum displacement among all is the maximum cumulative motion of the chosen point.

2.2. Objects

In order to extract the features like *# of objects in motion*, *# of macroblocks in motion* and *presence of object at the border of the image* we compute the local motion present in the video. The local motion is computed as the translational motion of the macroblocks. However, the local motion needs to be computed with respect to the global motion in the video. Thus it is subtracted from the global motion. The motion for macroblocks is searched within the vicinity of [-16, 16] with respect to the center of the macroblock. The size of a macroblock is 16 x 16. We estimate the following features based on the local motion:

2.2.1. Number of objects in motion (N)

After identifying the macroblocks that indicate local motion; since the local motion estimation is not accurate we eliminate the macroblocks that do not have any neighboring macroblock having the motion. The number of the objects is estimated by region growing algorithm from macroblocks having motion. Note that we do not require an accurate estimation of the number of objects.

2.2.2. Presence of objects at the borders of a frame (P)

We also determine whether an object appears at the borders of a frame or not. This helps us identify the type of video. Especially, in tracking videos, the objects are maintained close to the center of a frame.

2.2.3 Number of macroblocks in motion (M)

This feature counts the number of macroblocks that have motion. The number of objects might be misleading if there are multiple moving objects whose have a neighboring macroblock with another object. Thus *number of macroblocks* in motion helps to increase the accuracy of the same. All these features are identified for each frame in the video.

3. CLASSIFICATION FOR SPRITE GENERATION

We have determined 6 classes in Table 1 and pre-determined a decision tree for classifying these videos. At the first level of the tree, the presence of global motion is checked. At the second level, the moving objects are evaluated. The moving object evaluation includes determining the number of objects in motion in each frame determining the number of macroblocks in motion for each frame, and detecting the presence of moving objects at the border of the frame. The 6 classes are as shown in Figure 4.

We determine the significance of the global motion with respect to translational parameters. Motion less than 2 pixels is regarded as no motion whereas motion more than $\lfloor \max(w,h)/10 \rfloor$ is regarded as significant motion. Any GM between 2 to $\lfloor \max(w,h)/10 \rfloor$ pixels is regarded as minimum GM.

If the number of objects is more than 2, multiple moving objects are assumed. In case of minimum global motion, at the second level we check if there is any fixed pattern for the global motion that exists throughout the video.

Class Name	Class Details
Static Video (SV)	No global motion (GM), no moving objects (MOs) and no macroblocks(MBs) in motion
News, Educational and Surveillance Video (NES)	No GM, with >= 2 MOs and MBs in motion
Earthquake Video (EV)	Minimum GM (between 2-17 pixels), with fixed pattern in GM
Commercial Video (CV)	Minimum GM (between 2-17 pixels), without fixed pattern in GM
Scenery and sports Video (SSV)	Significant GM (>17 pixels), with < 2 MOs MBs in motion
Complex Video (CoV)	Significant GM (> 17 pixels), with >2 MOs and numerous MBs in motion

Table 1: Class Identification Details

Based on the presence of the pattern, the video is classified into the earthquake or commercial video class. Figure 5 provides the horizontal and vertical motion parameters for an earthquake video. The commercial videos do not have a pattern as earthquake videos have.

Figure 3: Dataset used for experiments

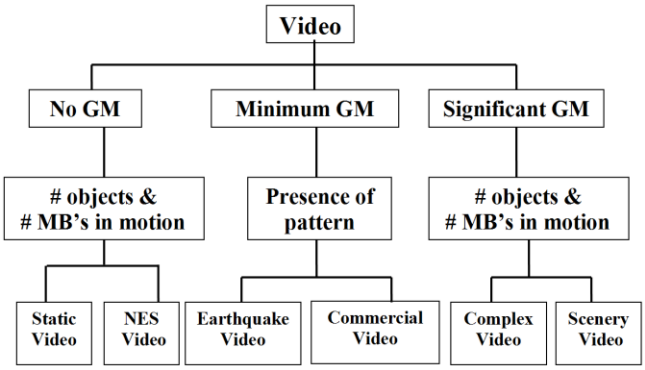

Figure 4: Classification Tree

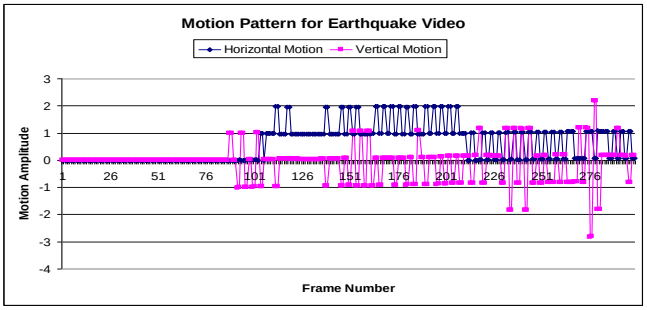

Figure 5: Motion pattern for earthquake video

4. EXPERIMENT RESULTS

Our data set includes a variety of videos such as educational, news, travel, entertainment, sports, animation, and documentary videos. From our collection of 100s video of varying length, and having variety of features, we selected 28 different videos for the experiments. These test sequences are split into shorter sequences and these short sequences are used in the experiments to study the effect of varying the number of frames being processed. The duration of the video is decided based on the number of frames processed for the video. The duration feature can be used for an available video rather than for live encoding of the video. The sprite coding may not yield efficient coding if the duration of a video shot is very short (e.g., several seconds). If the duration of the shot is known to be short, the sprite coding can be avoided since a short scene may not include significant global motion. The duration of the video is also important in the accuracy of extracting features regarding objects in the video.

The frame rate of our videos is 30fps and the size of these images is 176 X 144 pixels. Figure 3 shows the last frame of the selected 28 videos from our dataset. Out of these 28 videos 2 belong to static class, 9 to the NES class, 1 to the earthquake class, 5 to the commercial class, 2 to the scenery class and 8 belong to the complex video class.

4 sample different videos from our data set are provided in Figures 8. The first set is from the Chinese music -2 video; this video belongs to the static class and has no global or local motion in it. The second set is from the Earthquake where we observe slight vibrational motion pattern along with the local motion, which consists of a person walking and a motor vehicle passing on the road. The vertical and horizontal motion pattern for this video is as shown in Figure 5. The next set of frames (Figure 8) is from a scenery video that provides the panoramic view of a sand dune. This video has a constant GM and no LM. The final set of frames is from a tracking video in which local motion of a set of people in the frame is tracked.

The following table provides the classification results from our approach for few videos in our dataset. In this table AC stands for actual class, PC all refers to the predicted class when all frames were processed, PC 150 refers to 150 frames were processed, PC 75 refers to 75 frames were processed and PC 30 refers to 30 frames were processed. Figure 7 shows the accuracy obtained on varying the number of frames being processed for the video.

Video Title	AC	PC all	PC 150	PC 75	PC 30
Ayna Australia	CV	CV	CV	CV	CV
Ayna India – 1	NES	NES	NES	NES	NES
Global motion - 1	CV	CV	CV	CV	CV
Ayna India – 2	CoV	SSV	SSV	CV	CV
Bridges Of Dialogue	NES	NES	NES	NES	SV
Coast Guard	CoV	CoV	CoV	CoV	CV
Chinese Music - 1	SV	SV	SV	SV	SV
Chinese Music – 2	SV	SV	SV	SV	SV
Earthquake	EV	EV	EV	EV	EV
Exploring Turkey – 2	SSV	SSV	SSV	SSV	SSV
Panoramic View	SSV	SSV	SSV	CV	SSV

Table 2: Table showing sample results

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a method of classifying videos for sprite generation based on the motion parameters. In our results we have achieved 82% accuracy for the classifying videos for 6 video classes. We also used the DTREG predictive modeling software to generate the decision tree and obtained accuracy of 79%. Our previous approach achieved promising accuracy of 70% which was a good accuracy as a starting point. Compared to our previous approach, the improved approach analyzes the effect of varying number of frames processed to determine the class of the class of the video. The accuracy obtained after reducing the number of frames from all to 150 is close to the accuracy obtained after processing all the frames.

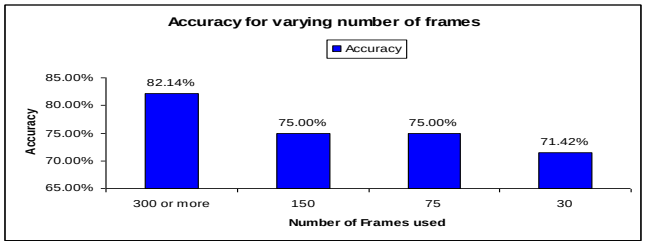

Figure 7: Accuracy chart

The accuracy drops marginally when number of frames is reduced to 75 and 30. This shows that it is possible to reduce the number of frames processed before classifying a video suitable for sprite generation, thus reducing the number of computations and in turn the processing time of the video. As future work, we plan to a) increase the number of videos for experiments, and b) implement the pattern detection in case of significant global motion as early as possible.

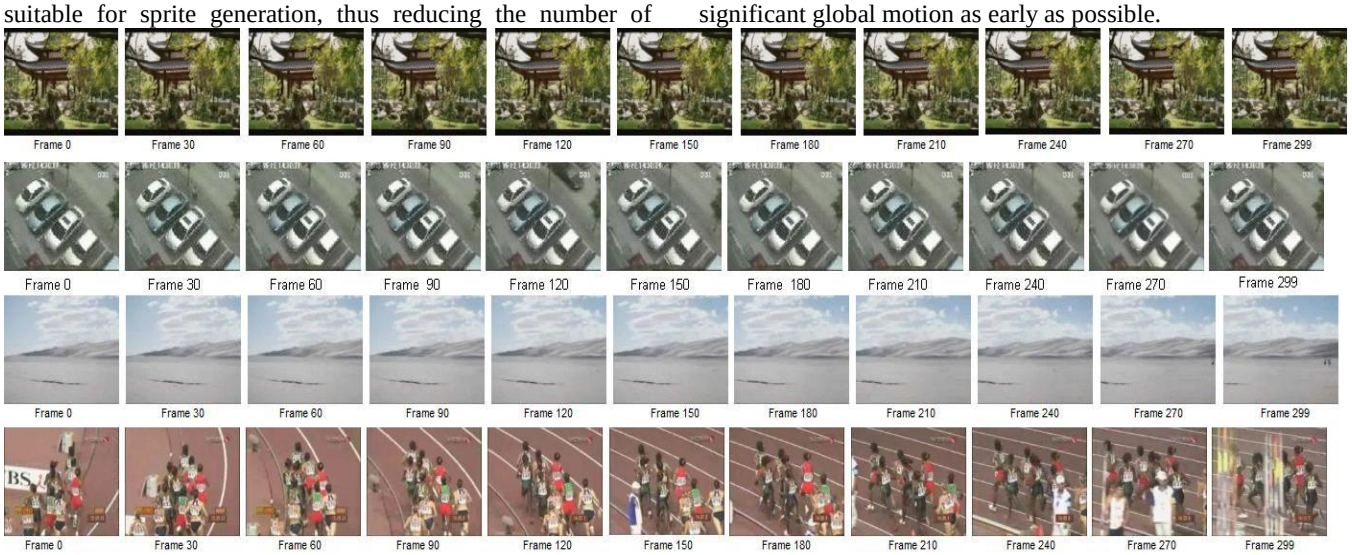

Figure 8: Frames for Chinese Music – 2, Earthquake, Panoramic and Tracking Videos

6. REFERENCES

[1] Gibert, X.; Huiping Li; Doermann, D., "Sports video classification using HMMS," *Int. Conference on Multimedia and Expo*, vol.2, no., pp. II-345-8 vol.2, 6-9 July 2003.

[2] L.Q. Xu and Y. Li, "Video classification using spatial-temporal features and pca", Proceedings ICME Multimedia and Expo, volume 3, pp. 485-8, 2003.

[3] Ye Yuan; Jun-Yi Shen; Qin-Bao Song, "A new rule-based video classification approach," *Machine Learning and Cybernetics, 2003 International Conference on* , vol.1, no., pp. 225-230 Vol.1, 2-5 Nov. 2003

[4] Papadopoulos, Georgios Th.; Mezaris, Vasileios; Kompatsiaris, Ioannis; Strintzis, Michael G., "Estimation and representation of accumulated motion characteristics for semantic event detection," *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on* , vol., no., pp.41-44, 12-15 Oct. 2008

[5] Glasberg, R.; Schmiedeke, S.; Kelm, P.; Sikora, T., "An automatic system for real-time video-genres detection using high-level-descriptors and a set of classifiers," *IEEE International Symposium on Consumer Electronics,* vol., no., pp.1-4, 14-16 April 2008

[6] Shi, Xiangqiong; Schonfeld, Dan, "Video Classification and Mining Based on Statistical Methods for Cross-Correlation Analysis," *Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on,* vol., no., pp.586-590, 26-29 Aug. 2007.

[7] Wen-Nung Lie; Chen-Kang Su, "News video classification based on multi-modal information fusion," *Image Processing, 2005. ICIP 2005. IEEE International Conference on* , vol.1, no., pp. I-1213-16, 11-14 Sept. 2005

[8] R. S. Aygun and A. Zhang. "Reducing Blurring-Effect in High Resolution Mosaic Generation" *2002 IEEE International Conference on Multimedia and Expo, Lausanne, Switzerland, 2002,* Vol. 2, pp. 537-540.

[9] http://www.chiariglione.org/mpeg/standards/MPEG-4/MPEG-4.htm

[10] Wiegand, T.; Sullivan, G.J.; Bjontegaard, G.; Luthra, A., "Overview of the H.264/AVC video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on* , vol.13, no.7, pp.560-576, July 2003

[11] Abhidnya Deshpande and Ramazan Savas Aygun. "Video classification for sprite generation in dynamic video encoder" unpublished to *2009 IEEE International Conference on Image Processing*, Cairo, Egypt. Nov 7-11, 2009.

[12] G. Kuhne, S. Richter, and M. Berer. "Motion-based segmentation and contour-based classification of video objects", *In Proc ACM MM Canada 2001*.

[13] Yi Haoran; Rajan, D.; Chia Liang-Tien, "An efficient video classification system based on HMM in compressed domain," *Proceedings of the 4th Pacific Rim Conference on Multimedia*, vol.3, no., pp. 1546-1550 vol.3, 15-18 Dec. 2003

[14] "Iso/iec 14496-10:2003, information technology: Coding of audio visual objects – part 2, also itu-t recommendation h.264 advanced video coding for generic audio visual services".

[15] Long Thang To, "Video Object Segmentation using Phase-based Detection of Moving Object Boundaries", Ph.D. thesis, University of New South Wales, 2005.

[16] http://www.dtreg.com/, accessed on May 29, 2009