A Tool For Query and Analysis of MPEG Encoded Video

Chris Kreucher and Sridhar Lakshmanan[♦]

Abstract — **This paper presents a tool for efficient query and analysis of MPEG encoded video. The image sequences for the video are obtained via a vehicle-mounted forward-looking camera. The tool accepts queries that signify both temporal and geometric events as the vehicle traverses common roadways, such as "identify portions of the MPEG encoded video where the vehicle is making a lane change maneuver," or "identify portions of the video where the vehicle is going around a tight curve," etc. Both query and analysis are done directly in the encoded domain.**

I. Introduction

The enormous amount of image and video data that typifies many modern multimedia applications mandates the use of encoding techniques for their efficient storage and transmission. The video encoding (compression) standard of choice in many new personal computers, video games, digital video recorders/players /disks, digital television, etc. is the one adopted by the Motion Pictures Expert Group (MPEG) [1]. Since this standard is being so widely accepted, it is important to develop tools that will enable MPEG encoded videos to be easily manipulated. These include tools that directly manipulate MPEG encoded video to achieve commonly desired functions such as overlap, translation, scaling, linear filtering, rotation, pixel multiplication, etc. - see [2].

While such tools are very useful for certain types of popular video editing tasks, they are not so relevant for other important functions such as feature extraction, query, browsing etc. Many techniques do exist for feature extraction, query and browsing of spatial domain video – see [3-5] for several different examples. However, they all involve the expensive and inefficient operation of decoding MPEG encoded videos before they can be applied.

This paper presents a tool for shape-based feature extraction, query and browsing of MPEG encoded video without the need for any expensive/inefficient decoding:

- The shape of the lane/pavement markers present in the I-frames is estimated using a Bayesian inference technique. A set of DCT-based lane edge features is used to arrive at a likelihood probability. A global appearance model for the shape of lanes in the image plane is derived, and this model is used to obtain a prior probability. The two probabilities are combined together using Bayes' rule, and the lane shape estimation problem is re-formulated as a posterior probability maximization problem. A multi-resolution optimization algorithm is used to perform this maximization.
- The lane shape process in P-frames is a little different. The motion vectors of the P-frame macro-pixel-blocks are used to identify which of them came from previous lane containing I or P-frame macro-pixel-blocks. P-frame macro-blocks that indeed came from previous lane containing I or P-frame macro-blocks are used to obtain an initial guess of the shape of the lane in this new P-frame by a non-linear least squares parameter fit. This initial guess is subsequently refined by a local maximization of the new P-frame posterior probability with respect to the lane shape parameters.
- If the encoded video stream contains intermittent B-frames, then lanes are identified in these B-frames by a procedure very similar to that of P-frames. The only difference is that the initial guess of the lane's shape in this B-frame is based on the estimates of lane geometry from both past and future frames.
- The lane's geometry as it evolves from frame-to-frame is finally analyzed. If this evolution indicates a geometrically and/or temporally relevant event that correlates with the query in question, then the corresponding portions of the encoded MPEG video clip are tagged. For example, if the query in question is "identify portions of the MPEG encoded sequence where the vehicle is making a lane change maneuver," then the frame-to-frame evolution of the lane offset parameters is closely scrutinized and if a trend corresponding to a lane change is detected then those portions of the video are tagged as matching the query.

II. Lane Shape Estimation in I-Frames

The lane shape estimation process in such I-frames consists of three steps:

- First, lane edge features are extracted in the encoded domain by using a small set of DCT coefficients. The spatial gradient field's magnitude and orientation are both nicely captured by these DCT coefficients. Using these lane edge features, a likelihood probability is constructed over the observation space.
- Second, a model of the shape of lane edges in the image plane is derived. Using this model, a deformable

[♦] Department of Electrical and Computer Engineering. The University of Michigan-Dearborn, 4901 Evergreen Road, Dearborn, MI 48128-1491 USA

template based prior probability is constructed over the lane shape parameter space.

- Finally, by combining the prior and likelihood probabilities using Bayes' rule, the lane shape estimation problem for I-frames is reformulated as a posterior probability maximization problem. In other words, the lane shape estimation problem is reduced to finding the global maximum of a four-variable function. A multi-resolution search algorithm is used to obtain this maximum.

Step 1-Edges are the feature of interest for estimating the shape of lanes, and edge orientation and strength are indeed very important in delineating pixels that could possibly lie on lane and pavement markers from pixels that are irrelevant. Specifically, circularly concentric lane markers in the ground will assume "diagonally dominant" orientations in the focal plane due to perspective transformation inherent in the image sensing process. To detect such diagonally dominant edge features in the I-frames, the tool uses a technique similar to [6] as its first step. For each block of 8×8 pixels, the amount of diagonally dominant edge energy contained in that block is decided by examining a special set of 12 DCT coefficients – see Figure 1. Figure 2 shows several examples of how effective these 12 special bases are from the standpoint of lane shape estimation. For each of the original images in Figure 2, the corresponding feature images are obtained by just summing the squares of its 12 special DCT decompositions. As one can see, in each case, despite the original image having features/edges of various strengths and orientations, the corresponding DCT feature images contain only information about those edges that are diagonally dominant.

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 1. The matrix that represents which 12 of the 64 bases capture diagonally dominant edges.

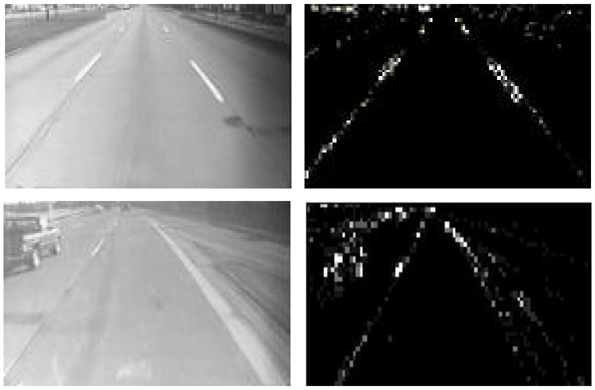


Figure 2. Effectiveness of the diagonal edge energy.

Step 2-The algorithm uses a global shape model to predict the manner in which lane markers appear in the image plane. As commonly done [7], this paper also assumes that lane markers are circular arcs on a flat ground plane. For small-to-moderate curvatures, a circular arc with curvature *k* can be closely approximated by a parabola of the form:

$$x = \frac{1}{2} k\, y^2 + m\, y + b \tag{1}$$

These parabolas in the ground plane map to hyperbolae in the image plane – see [7] for a derivation.

$$c = \frac{k'}{r} + b'\, r + vp \tag{2}$$

For left and right lane edges defined by concentric arcs, the approximation is made that the arcs have equal curvature and equal tangential orientation where they intersect the *X* axis, so *k'* and *vp* will be equal for the left and right lane edges. As a result, the lane shape in an image can be defined by four parameters k', b'_{LEFT}, b'_{RIGHT} and vp -k' is linearly proportional to the curvature of the arc on the ground plane, *vp* is a function of the tangential orientation of the arc on the ground plane, b'_{LEFT} and b'_{RIGHT} are functions of the offset of the arc from the camera on the ground plane.

Step 3-Real world lanes are never too narrow, wide or curved. A prior probability density function (pdf) is constructed over the lane shape parameters k', b'_{LEFT}, b'_{RIGHT} and *vp*:

$$\begin{aligned} \mathrm{P}(k', b'_{LEFT}, b'_{RIGHT}, vp) \propto \{ & \mathrm{atan}\, a[b'_{RIGHT} - b'_{LEFT} - 1] - \\ & \mathrm{atan}\, a[b'_{RIGHT} - b'_{LEFT} - 3]\} \times \{1 - b\left(\frac{k'}{x}\right)^2\}, \end{aligned} \tag{3}$$

where $a = 10, b = 0.01$ and $c = 600$ values are chosen to reflect the *a priori* knowledge. It is assumed that given the values of k', b'_{LEFT}, b'_{RIGHT} and *vp*, the probability of the observed image having the DCT feature values (the ones described previously) is given by the likelihood pdf:

$$\begin{aligned} & P(DCT\ feature\ values \mid k', b'_{LEFT}, b'_{RIGHT}, vp) \\ & \propto \sum_{i,j} \sum_{k,l \in Ci,j} (dct\_coeff(k,l))^2 \end{aligned} \tag{4}$$

These two pdfs are combined using Bayes' rule, and the lane shape estimation problem is reduced to one of finding the global maximum of a posterior pdf. The MAP (maximum *a posteriori*) estimate is found by a coarse-to-fine search over the four-dimensional parameter space of k', b'_{LEFT}, b'_{RIGHT} and *vp*.

III. Lane Shape Estimation in P and B Frames

The MPEG encoding standard uses inter-coded (P and B) frames to substantially improve the compression ratio. Information regarding the motion of macro-pixel-blocks in a frame and the difference in DCT coefficients between motion-related macro-pixel-blocks is used to reduce the number of bits necessary to encode that frame. This technique is useful in efficient encoding of image sequences, because the scene changes little from frame-to-frame. By taking advantage of the motion information present in P- and B-frames, our tool estimates the lane shape more quickly in the P- and B-frames than it does in the I-frames.

For P-frames, MPEG breaks the frame into 16x16 macro-pixel-blocks and encodes the DCT of these blocks in terms of a motion vector and an error vector. The motion vector contains information indicating what past macro-pixel-block that new macro-pixel-block is being encoded with respect to, i.e., where this new macro-pixel-block came from. The error vector, on the other hand, represents the change in the DCT coefficient values for the new block relative to the old block. In cases where encoding in this fashion would require more bits than just a standard (I-frame type) encoding of this new block, the new block is intra-coded just like an I-frame block, and no motion or error vectors are present.

Lane shape estimation in P-frames[1] is a three-step process as well, only the actual steps are a little different than in I-frames:

- First, the P-frame motion vectors are used to approximate the motion of the lane from frame-to-frame, without any regard of the error vector.
- Second, using this approximate lane motion, a non-linear least squares fit is performed to make an initial guess as to what the exact shape of the lanes is in this new P-frame.
- Finally, a local maximization of the P-frame posterior pdf (derived using the error vectors) is performed to refine the initial guess.

Step 1-Each 16x16 macro-pixel-block of the P-frame has a motion vector indicating its origin in the preceding I- or P-frame. Since that previous frame has already been processed, the shape of the lanes in that frame is known. If a P-frame macro-pixel-block is deemed to have originated from a lane containing macro-pixel-block in the past P- or I-frame, that new P-frame block is so marked. A distinction is made between those blocks that came from left lanes and those that came from right lanes. By inspecting all macro-pixel-blocks in the current P-frame, a tertiary map indicating how the previously estimated lane has moved from the past frame to the current frame is generated.

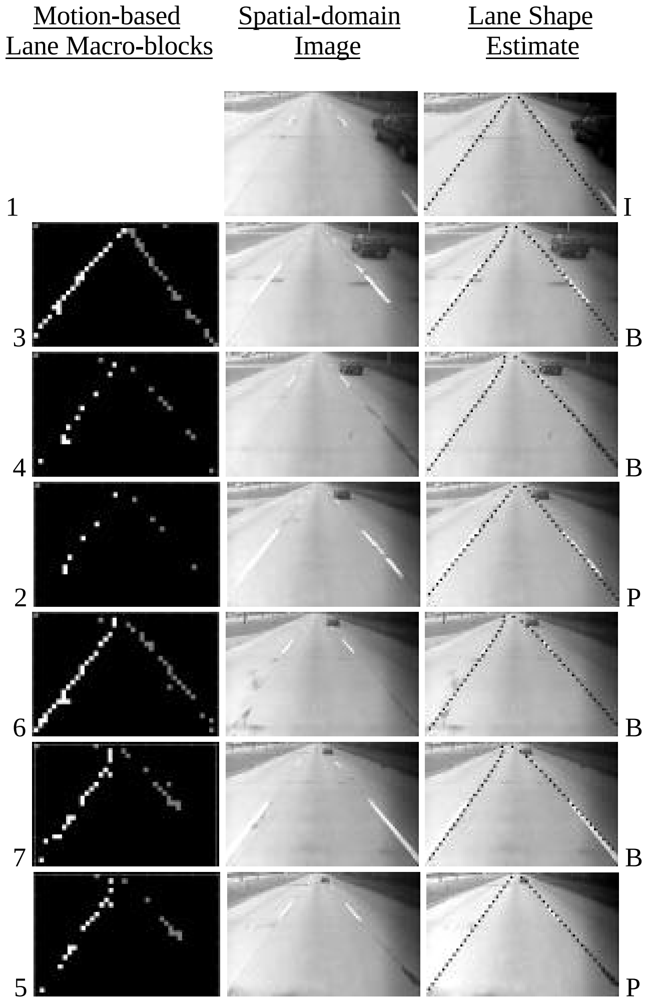
Step 2-Using the tertiary map as data samples of the non-linear regression given by eq. (2), a least-squares estimate of the regression parameters k', b'_{LEFT}, b'_{RIGHT} and vp is obtained from that data. This non-linear least squares estimate is our tool's initial guess as to what the lane's shape is in this new P-frame. There are two important reasons why this easy to calculate initial estimate is not sufficiently accurate and requires a proper refinement before it can be accepted as the final lane shape estimate for the new P-frame:

- First, since the initial estimate is based on motion vectors alone, those blocks that are inter-coded (have no associated motion vector) are ignored. This includes blocks that correspond to areas of the current P-frame that have newly appeared, and are not present in the previous image. Also included are macro-pixel-blocks that are significantly different than the past due to lighting, shadowing, occlusion etc.
- Second, some present macro-pixel-blocks that are encoded with respect to past lane containing macro-pixel-blocks do not actually contain lanes. Typically, these are blocks that have lane-like properties in terms of the strength and orientation of the edges they contain.

Step 3-Using the method discussed in section II, a feature image for the new P-frame is computed. A posterior pdf over the lane shape parameter space is then formulated for this new frame as in section II. The initial guess is used to initiate a simple locally exhaustive search of the posterior pdf over a very small area of the lane shape parameter space. This local search usually results in a good refinement of the initial guess, and is accepted as the final estimate of the lane's shape in the new P-frame.

IV. Experimental Results

The lane shape estimation processes of sections II and III, and how they fit together is illustrated first in Figure 3 on a 10 frame MPEG encoded video stream.



[1] The extension of this procedure to lane shape estimation in B-frames is conceptually simple. Unlike P-frames, that encode current frames with respect to preceding frames only, the encoder allows B-frames to use information from both past and future frames. Hence, the motion vectors in the new frame will refer to both past and previous frames.

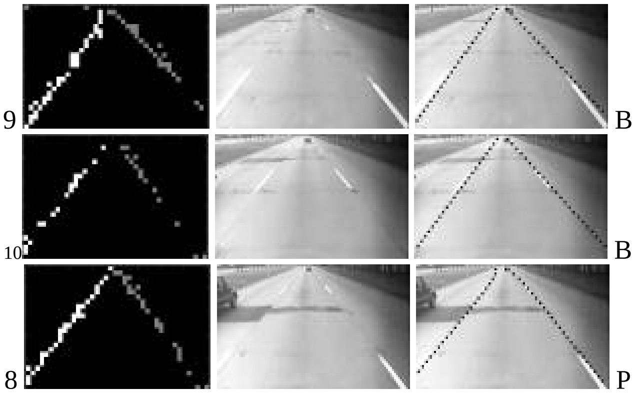


Figure 3. Lane shape estimation in an MPEG encoded video stream. In each row, the number on the extreme left indicates processing order, and the letter on the extreme right indicates frame type.

This lane shape estimation process is subsequently used to query MPEG encoded videos of common roadway scenes. The results of two such queries, one to locate a temporally meaningful event and another to locate a geometrically meaningful one are presented in figures 4 and 5 – see valhalla.umd.umich.edu/~ckreuche for additional results . The query in figure 4, namely, "find portions of the video where the vehicle is making a lane change maneuver," is translated to mean find portions of the video where either of the two lane offset parameters b'_{LEFT} or b'_{RIGHT} undergoes a smooth zero crossing – see reference [8] for details. The query in figure 5, namely, "find portions of the video where the vehicle is going around a tight curve," is translated to mean find portions of the video where the curvature parameter k' continues to have a magnitude much larger than zero over a series of frames – see reference [9] for details.

REFERENCES

[1] Haskell, Barry G., *Digital video: an introduction to MPEG-2,* Chapman & Hall, New York 1997.

[2] Chang, Shih-Fu and Messerschmitt, David G., "Manipulation and compositing of MC-DCT compressed video," *IEEE Journal on Selected Areas in Communications,* vol. 13, pp. 1-11, 1995.

[3] Ahmad, T., Taylor, C.J., Lanitis, A., and Cootes, T.F., "Tracking and recognizing hand gestures using statistical shape models," *Image and Vision Computing*, vol. 15 pp. 345-352, 1997.

[4] Yang, Jie and Waibel, Alex, "Real-time face tracker," *IEEE Workshop on Applications of Computer Vision,* pp. 142-147, 1996.

[5] Zhong, Yu, *Object Matching using Deformable Templates,* Ph. D. Thesis, Michigan State University, 1997.

[6] Y. S. Ho and A. Gersho, "Classified transform coding of images using vector quantization," *IEEE International Conference on ASSP,* pp. 1890-93, 1989.

[7] Kluge, K.C. and Lakshmanan, S., "A deformable template approach to lane detection," *Proceedings of the IEEE Intelligent Vehicles Symposium,* pp. 54-59, 1995.

[6] Yeo, Boon-Lock Liu, Bede, "Visual content highlighting via automatic extraction of embedded captions on MPEG compressed video," *Proceedings of SPIE - The International Society for Optical Engineering* vol. 2668, pp. 38-47, 1996.

[7] Wang, Hualu and Chang, Shih-Fu, "Highly efficient system for automatic face region detection in MPEG video," *IEEE Transactions on Circuits and Systems for Video Technology,* pp. 615-628, 1997.

[8] Kluge, K.C., Kreucher, C.M., Lakshmanan, S., "Tracking Lane and Pavement Edges Using Deformable Templates," *Proceedings of SPIE*, 1998.

[9] Kreucher, C.M. and Lakshmanan, S., "POIROT: A System for Querying JPEG Encoded Image Databases", under review *IEEE Transactions on Image Processing*, 1998.

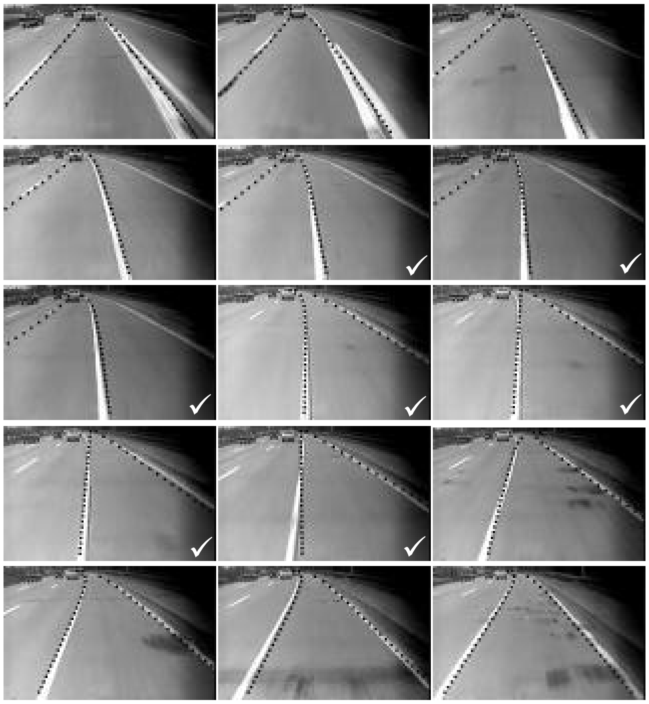


Figure 4. Results of querying the MPEG stream with "find portions of the video where the vehicle is making a lane change maneuver." Retrieved images are ticked.

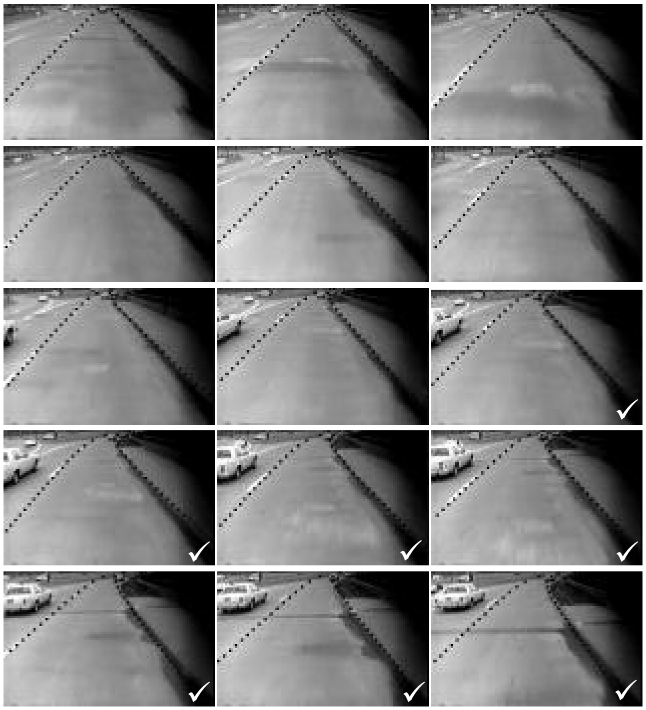


Figure 5. Results of querying the MPEG stream with "find portions of the video where the vehicle is going around a tight curve." Retrieved images are ticked.