REAL-TIME TEMPORAL TEXTURE CHARACTERISATION USING BLOCK-BASED MOTION CO-OCCURRENCE STATISTICS

Ashfaqur Rahman and Manzur Murshed

Gippsland School of Comp. & IT, Monash University, Churchill Vic 3842, Australia
{Ashfaqur.Rahman, Manzur.Murshed}@infotech.monash.edu.au

ABSTRACT

Contemporary temporal texture classification methods use pixel based features thus making the process slow for time sensitive applications like video indexing and surveillance. In this paper, a real-time classification technique is presented by using readily available block based motion vectors. Experimental results demonstrate the ability of the proposed technique to classify a large set of temporal textures in real-time with high accuracy.

1. INTRODUCTION

A large class of objects commonly experienced in real world scenario exhibits characteristic motion with indeterminate spatial and temporal extent. The motion assembly by a flock of flying birds, water streams, fluttering leaves, and waving flags are some of the most common examples that serve to illustrate such motion. Contemporary literature [1] coined the term *temporal texture* to identify collectively such motion patterns that exhibit spatiotemporal regularity but have indeterminate spatial and temporal extent.

The phenomena commonly observed in temporal textures, together with the vast domain in which they exist, has prompted many researchers [1]–[9] to formulate techniques to classify, recognize, and synthesize the distinctive motion patterns. Polana and Nelson [1] first used an approximated motion measure (normal flow) to recognise temporal textures using statistical features obtained from successive frame pairs. By criticising this work for lacking any mechanism to handle temporal evolution, Bouthemy and Fablet [3] used statistical features on the temporal motion distribution. However, classifying spatiotemporal regularities using either space or time domain information, while ignoring the other domain, has serious flaws.

To address this, Otsuka *et al.* [4] used features based on the dominant spatiotemporal motion trajectory surface. However, it is highly unlikely for temporal textures to have a dominant trajectory surface with indeterminate spatial and temporal extent. Even if such a surface exists, extraction of the surface is not an easy task and its accuracy is questionable [6]. Peh and Cheong [6] obtained statistical features spatially only after superimposing consecutive frames to capture time-integral property of the texture. This idea is, however, technically flawed on the ground that merging a long sequence inevitably looses a lot of characteristic information while it is obvious that in order to capture the time-integral property accurately, the sequence has to be long.

Recently, a number of researchers [7]–[9] developed temporal texture classification techniques by exploiting the spatiotemporal regularity property in a convincing manner. Saisan *et al.* [7] and Smith *et al.* [8] used only pixel intensity information and thus avoided any motion related measure by extracting features using an autoregressive model and 3D wavelets respectively. Fablet *et al.* [9] extended the notion of temporal motion distribution statistics by introducing temporal cliques to capture some extent of spatial motion regularity.

In this paper, a real time temporal texture classification technique is developed using spatiotemporal statistics on block-based motion vectors, readily available in MPEG-1/2/4 and H.26X image sequences. Although contemporary methods fail to maintain an explicit ratio between spatial and temporal domain features the proposed method utilizes the natural proportion between the two domains. Experimental results clearly demonstrate that the proposed technique is capable of classifying temporal textures with high accuracy.

2. MOTIVATION

Careful scrutiny of the existing techniques reveals the following two facts. First, all of the convincing pixel-based techniques [7]–[9] are not suitable for any real-time application e.g., video indexing and retrieval and real-time video surveillance, as the computational complexity for extracting features from the motion measures of all the pixels of an image sequence will be of order $cO(V)$ where V is the pixel volume of the sequence and c is a constant greater than 1. Second, all the previous works, except those in [7] and [8], already established that temporal textures can be classified using an approximated motion measure such as normal flow. The obvious approach to make the characterisation technique faster involves using a representative motion measure for a block of pixels instead of individual pixels provided the block measure is a good approximation.

Block-based motion vectors, readily available in MPEG-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

1/2/4 and H.26X image sequences, are calculated with a view to improve coding efficiency. However, they still represent some degree of true motion measure that is successfully exploited in motion indexing of block-based videos [11], motion-based video indexing and retrieval [12], and neighbouring motion vector prediction [13]. The approximation of true motion by the block based motion vectors of a temporal texture video is even stronger which can be explained from their spatiotemporal motion regularity. As temporal texture videos possess spatial motion uniformity, pixels of a small spatial block are supposed to undergo uniform displacement. It is, therefore, highly likely that the true displacement vector will also point to the block with the minimum error difference which is also sought by the full search [14] motion vector estimation technique. Using the block-based motion vectors for classifying temporal textures has an added benefit of encoding some degree of spatial correlation information to improve classification accuracy.

In this paper, a temporal texture classification technique is developed using spatiotemporal statistics on block-based motion vectors. From classification accuracy and implementation feasibility point of view, the pixel-based technique of Fablet *et al.* [9] has been a natural choice to extend the idea into block-based domain. The other promising ideas in [7] and [8] cannot be transformed easily as unique intensity for a block of pixels is not well defined. In order to form the basis of technical reasoning, the statistical motion modelling is presented in the next section in the light of the model in [9].

3. STATISTICAL MOTION MODELLING

Due to spatiotemporal regularity, if the volume of motion related quantities of a temporal texture video is divided into a sequence of parallel grids along any direction, some form of causal relationship will exist between successive grid pairs. Any such sequence of K grids $g = (g_k)_{k=0,\cdots,K}$ can be realised as a first-order Markov chain $G = (G_0, \ldots, G_K)$ such that

$$P(g) = P(g_0) \prod_{k=1}^{K} P(g_k | g_{k-1}), \tag{1}$$

where $P(g_0)$ represents the *a priori* distribution of the first grid in the sequence. In practice, however, $P(g_0)$ is assumed constant. In order to design purely causal model, a point $g_k(b)$ is assumed independent conditionally to corresponding point $g_{k-1}(b)$. Thus (1) can be rewritten as,

$$P(g) = P(g_0) \prod_{k=1}^{K} \prod_{\forall b} P(g_k(b) | g_{k-1}(b)). \tag{2}$$

Assuming an exponential distribution $P(g_k(b)|g_{k-1}(b)) \propto \exp \phi(g_k(b), g_{k-1}(b))$ (2) becomes

$$P(g) = \exp\left[\sum_{k=1}^{K} \sum_{\forall b} \phi(g_k(b), g_{k-1}(b)) \right]. \tag{3}$$

Let Γ be the co-occurrence matrix over the grid sequence obtained as follows:

$$\Gamma(p,q) = \sum_{k=1}^{K} \sum_{\forall b} \delta(g_k(b), p) \delta(g_{k-1}(b), q) \tag{4}$$

where

$$\delta(r,s) = \begin{cases} 1, & \text{if } r = s; \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The ϕ can then be calculated from this co-occurrence matrix as,

$$\phi(p,q) = \ln(P(p|q)) = \ln \frac{\Gamma(p,q)}{\sum_{\forall m} \Gamma(m,q)}, \tag{6}$$

and (3) can be rewritten as

$$P(g) = \exp(\Gamma \bullet \phi). \tag{7}$$

It is clearly evident from (7) that the causal relationship due to the spatiotemporal regularity in a temporal texture video can be captured by the co-occurrence statistics on a sequence of parallel grids over the entire volume of motion related quantities.

4. PROPOSED TECHNIQUE

The most obvious grid sequence of a temporal texture video consists of each video frame as a grid with each grid point represented by a normal flow of the corresponding pixel or a block motion vector of the corresponding macroblock. Fablet *et al.* [9] used this conventional pixel-level grid sequence along with a clique structure ($\eta = \{1,5,9\}$ neighbouring points in the previous grid) for each point of a grid. Instead of calculating only one instance, η instances of Γ and ϕ matrices were calculated independently along each of the η neighbourhood directions in the clique and then some degree of optimality is achieved using a conjugate gradient method.

The clique structure was introduced with a view to incorporating spatial domain information in order to improve an earlier technique in [3], which used temporal co-occurrence only. The clique structure in a grid of blocks will fail to serve the purpose as it would then cover a large spatial area beyond any spatial regularity constraint. It can also be argued that indirect encoding of spatial domain information through cliques makes it very difficult to control an appropriate weight distribution between space and time domain information.

To overcome these problems, the proposed technique uses three instances of co-occurrence statistics obtained independently on the magnitude of block-based motion vectors. To incorporate time domain information, a temporal co-occurrence matrix is obtained on the conventional grid sequence (along the time axis) and to incorporate space domain information, two spatial co-occurrence matrices are obtained on two grid sequences along the direction of *x*- and *y*-axes. In order to keep the natural 1:2 proportion between time and space domains, each of the three co-occurrence statistics are given equal weights.

Dissimilarity between temporal textures in terms of co-occurrence statistics can be measured using KL-divergence [9]

that is defined as

$$KL(\mu_2 \| \mu_1) = \int \ln \frac{\mu_1}{\mu_2} d\mu \,, \tag{8}$$

where μ_1 and μ_2 are two probability distributions. $KL(\mu_2 \| \mu_1)$ measures the amount of information lost when a probability distribution μ_2 replaces μ_1. Given a type of co-occurrence matrix (spatial or temporal) KL-divergence between two video clips v_1 and v_2 can be approximated from (7) and (8) as

$$KL(v_2 \| v_1) \approx [\phi_1 - \phi_2] \bullet \frac{\Gamma_1}{V} \,, \tag{9}$$

where V is the volume of the image sequence. The divergence measure is made commutative by using the average of $KL(v_2 \| v_1)$ and $KL(v_1 \| v_2)$. Distance between video clips is obtained by Euclidian summation of co-occurrence matrix distances.

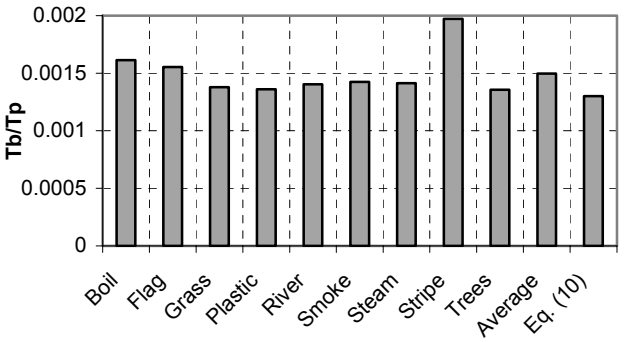

Fig. 1: Empirical time ratio of the proposed block-based method and the pixel-based spatiotemporal clique method in [9] for the nine temporal texture classes used in the experiments, their average, and the theoretical time ratio using (10).

5. COMPUTATIONAL COMPLEXITY

Without any loss of generality, the computational speed of the proposed technique is compared with that of the technique in [9] as the same of any other pixel-based technique will be of same order. Although pixel-based motion measures are not readily available in any of the existing popular video codecs, the computational complexity of deriving such measures are ignored in order to keep the comparison under equal footing. There are V and V/ab motion measures at pixel- and block-level respectively where V is the pixel volume of the sequence. Therefore, overall co-occurrence matrix calculation time for the pixel-based technique in [9], T_p, will be of order $\eta O(V)$ and the same for the proposed block-based technique, T_b, will be of order $3O(V/ab)$. Thus,

$$\frac{T_b}{T_p} = \frac{3O(V/ab)}{\eta O(V)} = \frac{(3/ab)O(V)}{\eta O(V)} = \frac{3}{\eta ab} \tag{10}$$

Fig. 1 shows T_b / T_p ratio for each class of image sequences used in the experiments for an example instance with block size of 16×16 and $\eta = 9$ where empirical results agree closely with the theoretical ratio of 0.0013.

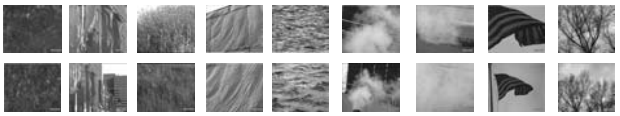

Fig. 2: Two representative video clips for the nine different temporal textures used in the experiments.

6. EXPERIMENTAL RESULTS

A series of nine sequences (Fig. 2) containing a representative set of different temporal textures, including boiling water, waving flag, and wind swept grass, was incorporated into the video database. For each texture type, two MPEG image sequences were partitioned into a total of 14 video clips (the exception being type C, where 12 clips were used) comprising twenty 352×240 pixel frames using 16×16 pixel macroblocks. To obtain true motion prediction some pre-processing steps are applied to motion vector frames - (i) A median filter is applied to eliminate noise; (ii) camera motion is compensated [10] for extracting true dynamic content. The magnitude domain of motion vector is constrained in the range $[0,20]$.

Conventional classification problems first tune cluster centres based on a training set and then use them to classify a test set. But for the proposed technique, features are not numeric values; rather they are composed of numeric entities. Thus such a tuning is not possible here. That's why classification is performed on the entire video database using *k*-NN classifier. A destination class for a video is decided based on *majority wins* rule. In case of a tie, the video is classified into an *undecided* group. The classification results are presented in Fig. 3 for the proposed method and spatiotemporal clique method for different values of *k*. Only odd values of *k* are used where *majority* is unarguably decidable. Although it is natural for any majority win *k*-NN classifier to degrade accuracy rate slowly with *k*, the degradation rate in Fig. 3 is significantly high due to the limited number of representative video clips per class.

In Fig. 3, classification accuracy of the spatiotemporal clique technique improves with η. This supports the underlying assumption in [9] that the clique structure captures some degree of spatial motion regularity. Although the improvement from $\eta = 5$ to $\eta = 9$ is insignificant, it is mainly due to the fact that they both represent the neighbourhood of Euclidian radius 1. However, using an arbitrarily high η is not a viable option as accuracy will improve only at the expense of computational speed, which is directly proportional to η as shown in (10).

Fig. 3 further reveals that the proposed technique, where temporal and spatial regularity are considered independently and then combined using their natural proportion, improves classification accuracy even further irrespective of whether it

is applied at block- or pixel-level. The proposed technique achieved on average 6% {(0.98–0.89)/0.89 ≈ 10%, (0.95–0.9)/0.9 ≈ 5.5%, and (0.91–0.88)/0.88 ≈ 3.5% for k = 1, 3, and 5} relative classification accuracy improvement over the best accuracy obtained by the spatiotemporal clique technique.

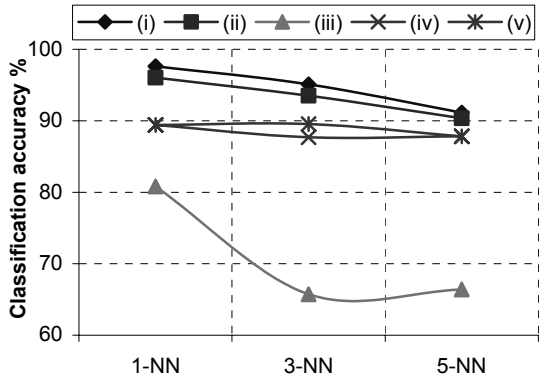

Fig. 3: Classification accuracy of the (i) proposed technique; (ii) proposed technique at pixel-level; (iii)–(v) spatiotemporal clique technique with $\eta = 1,5,9$ respectively; for different k-NN classifier.

The accuracy level of the proposed block-based technique is also relatively 1.5% better (across all k = 1, 3, and 5) than that of the technique applied at pixel-level. This improvement is due to the fact that when motion vectors are calculated they capture the spatial correlation of a block of pixels; whereas pixel based flows fail to capture such correlation. Due to space limitation, detailed classification results only for $k = 1$ is presented in Table 1 in terms of a confusion matrix where the proposed technique shows consistency in improving accuracy for all types of temporal textures with just one exception.

Table 1: Classification results for nine temporal textures. For each texture type, the first line (bold) refers to classification results using the proposed block based technique, and second (Italic) line refers to classification using spatiotemporal clique method ($\eta = 9$).

	A	B	C	D	E	F	G	H	I
A	**100** *100*								
B		**92.9** 92.9					**7.1** 7.1		
C			**100** *83.3*		*8.3*	*8.3*			
D				**85.7** *100*					**14.3**
E			*14.3*		**100** *85.7*				
F			*7.1*			**100** *85.7*	*7.1*		
G	*7.1*	*7.1*				*7.1*	**100** *71.4*	*7.1*	
H							*7.1*	**100** *92.9*	
I				*7.1*					**100** *92.9*

7. CONCLUSION

In this paper, a novel temporal texture classification technique has been presented where co-occurrence statistics of motion measures are obtained independently in spatial and temporal domains that are then combined with their natural 2:1 proportion in approximating the divergence among the image sequences. Although the proposed technique offers high accuracy level applied to both block- and pixel-level motion measures, the former can be realised in real time as it is 256 times faster than the latter for conventional block size of 16×16 pixel.

8. REFERENCES

[1] R. Polana and R. Nelson, "Recognition of motion using temporal texture," Proc. CVPR '92, pp. 129 -134, 1992.

[2] M. Szummer, and R. W. Picard, "temporal texture modelling," Int. Conf. on Image Processing, vol. 3, 1996.

[3] P. Bouthemy and R. Fablet, "Motion characterization from temporal cooccurences of local motion-based measures for video indexing," Int. Conf. on Pattern Recognition (ICPR'98), vol. 1, pp. 905-908 , 1998.

[4] K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii, "Feature extraction of temporal texture based on spatiotemporal motion trajectory," Proc. 14th Int. Conf. Pattern Recognition (ICPR'98), vol. 2, pp. 1047-1051, 1998.

[5] R. Fablet and P. Bouthemy, "Non parametric motion recognition using temporal multiscale Gibbs models," Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 501-508, 2001.

[6] C. -H. Peh, and L. -F. Cheong, "Synergizing spatial and temporal texture," IEEE Trans. on Image Processing, vol. 11, pp. 1179-1191, 2002.

[7] P. Saisan, G. Doretto, Y. Nian Wu, and S. Soatto, "Dynamic texture recognition," IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, vol. 2, 2001.

[8] J.R Smith, C. –Y. Lin, and M. Naphade, "Video texture indexing using spatio-temporal wavelets," Proc. IEEE international conference of Image Processing, Vol. 2, pp. 437-440, 2002.

[9] R. Fablet, P. Bouthemy, and P. Perez, "Nonparametric motion characterization using casual probabilistic models for video indexing and retrieval," IEEE Trans. on Image Processing, vol. 11, pp. 393 -407, 2002.

[10]G. Sorwar, M. Murshed and L. Dooley, "Fast global motion estimation using iterative least-square estimation technique," Fourth IEEE Pacific-Rim Int. Conf. on Multimedia (PCM-03), Singapore, 2003.

[11]E. Sahouria, and A. Zakhor, "Motion indexing of video," Proc. Int. Conf. on Image Processing (ICIP-97), vol. 2, pp. 526-529, 1997.

[12]Y. –F. Ma, and H. –J. Zhang, "Motion texture: a new motion based video representation," Proc. 16th Int. Conf. on Pattern Recognition, pp. 548-551, vol. 2, 2002.

[13]D.S. Turaga and C. Tsuhan, "Estimation and mode decision for spatially correlated motion sequences," IEEE Trans. on Circuits & Systems for Video Tech., vol.11, no.10, pp.1098-107, Oct. 2001.

[14]T. Sikora, "digital video-coding standards," IEEE signal processing magazine, Vol. 15, pp. 82-100, 1997.