Compressed Spatio-temporal Descriptors for Video Matching and Retrieval

Orkun Alatas Omar Javed Mubarak Shah

School of Computer Science
University of Central Florida, Orlando, FL 32816 USA
{alatas, ojaved, shah}@cs.ucf.edu

Abstract

The contents of a video can be described in terms of appearance and motion of the scenes. In this paper, we propose wavelet-based appearance descriptors and spline-based motion descriptors that can together represent the videos, and be used for matching and retrieval. The optical flows give the direction along which the gray levels have regular variations in time, and the wavelet decomposition can take advantage of this by decomposing the sequence along these directions, exploiting the redundancy. Such a compact representation of video is very crucial for storage and retrieval purposes. Therefore we show that these spatio-temporal descriptors can be used for 1) video retrieval and matching, and 2) calculating spatio-temporal similarity between articulated objects. The results are demonstrated on various kinds of sequences.

1. Introduction

With the advent of the digital age, there has been a rapid increase in the amount of video information available. With the growth in availability of video data the need to query and access the relevant data becomes critical. The videos have to be stored in a highly compressed form because of the huge storage requirements. Consequently any efficient video retrieval or matching algorithm has to use the compressed information directly.

Wavelet decomposition is one of the most common techniques used in still image compression. Video sequences can also be compressed by this method by decomposing groups of frames in three dimensions. In this paper, we present a warped wavelet decomposition that exploits the motion information in the sequence to decompose the video along the direction in which the video has regular variations. The motion information is estimated by a spline based representation of the local optical flow field.

The warped wavelet coefficients describe the motion compensated appearance of a group of frames and the spline coefficients describe the motion in this group. Thus the descriptor separates out and extracts both the significant appearance and motion features. Note that both information is useful for video retrieval systems. For example news videos usually have low motion content while the commercials have relatively high motion. Outdoor scenes might have a higher blue and green color content as compared to indoor scenes. We demonstrate that in a *query by example* video retrieval framework, relevant segments of videos can be accurately retrieved by performing a weighted matching of the motion and appearance coefficients.

The ability to people doing certain actions is important for surveillance related applications. We demonstrate that the spatio-temporal coefficients capture the gait of walking persons successfully. Given a few example of walking people, a correlation based measure is used to understand their action in a scene.

We will describe the related work in Section 2, the details of the spatio-temporal descriptor in Section 3. Recognition and video retrieval using the descriptors are discussed in Section 4. Finally, results are given in Section 5.

2. Related Work

In still image compression, Mallat and Le Pennec described a new scheme in [7], where they first found the lines of least intensity change in an image, then they warped the wavelet basis functions to decompose the image along these lines. Finally they made a change of basis, which was called *bandeletization*, to take more advantage of this directionality. This step drastically decreased the number of significant coefficients in the decomposition.

In video coding, 3D wavelets have been commonly studied in the past. Moyano et al showed an efficient 3D wavelet decomposition in [5]. They divided a video sequence into groups of frames and decomposed each of them in spatial and temporal dimensions.

Later Taubman described a more efficient approach in [3], where he registered the frames before decomposing them. The author first computed the optical flows using a deformable mesh model, then aligned the frames in time before decomposing them.

A large amount of research has been conducted in the field of video retrieval and matching. The video retrieval procedure presented by A. Del Bimbo et. al [2] required the segmentation and tracking of regions in a scene. The authors computed a Fisher distance measure between color regions in a sequence of frames. The wavelet decomposition of this measure was used as a descriptor of the sequence. E. Ardizonne et. al [4] matched videos using the color and motion direction histograms.

Fablet and Bouthemy [6] proposed a statistical approach for motion based object indexing and retrieval. The local motion of polygon regions marked by users is extracted. A kernel density estimator is used to model the motion of the marked object. The Kullback-Leibler divergence is used as a similarity measure for retrieval. Efros et. al [1] used a smoothed and half wave rectified motion descriptor for recognizing action of objects that were far from the camera, i.e. at resolutions where an object is around 30×30 pixels. A version of normalized correlation was used to compute the similarity of actions.

3 Computing the Compressed Descriptors

3.1 Compression with 3D Wavelets

In traditional 3D wavelet decomposition of video sequences, the wavelet basis functions are applied along the temporal and spatial dimensions over a group of frames (GOF). Although such a decomposition leads to good compression rates, there is still room for improvement by further exploiting the motion information in the video. This information lies in the optical flow, which tells us how the gray levels are displaced in time. If studied carefully, it may reveal the paths along which the temporal gradient is small. If the video can be decomposed along these paths, then the sequence can be represented by a smaller number of coefficients. A similar work was done by Taubman et al in [3], however their motion estimation was based on deformable mesh models and was inherently insufficient to provide smooth optical flow enough to achieve a good alignment of the pixels. We use a modified version of the regularization approach in [7], with the optical flows computed using splines. This model results in smoother fields of optical flow and increases the regularity in the alignment process.

3.2 Computing the Optical Flow

The spline-based motion model was first used by Szeliski et al in [8]. In this technique, 2D splines that are controlled by a small number of control points are used to approximate the flow. The control points are homogenously spread on the image grid, and each of them has horizontal and vertical displacement values $(\hat{u_j}, \hat{v_j})$. These values are solved by minimizing (1) using a preconditioned gradient descent method.

$$E(u,v) = \sum_i (I_t(x_i+u_i, y_i+v_i) - I_{t-1}(x_i,y_i))^2 \quad (1)$$

The dense optical flow (u_i, v_i) is a weighted linear combination of these values (2), which is highly correlated in neighboring pixels.

$$(u_i, v_i) = \sum_j (\hat{u_j}, \hat{v_j}) B_j(x_i, y_i) \quad (2)$$

where $B_j(x_i, y_i)$ is a bilinear spline function centered at a control point (x_j, y_j).

Representing the dense optical flow with a small number of control points is a desired property for matching and compression applications. Therefore the displacement values can be considered as *Compressed Motion Descriptors*, since all the motion information can be recovered from them.

3.3 Warping the Wavelet Basis

In a GOF, the optical flows form a 3D vector field, $(u(x,y,t), v(x,y,t))$. Such a representation allows us to introduce the concept of an *optical curve*, which can be defined as the integral curve of the optical flow (3).

$$U_i = \sum_t u(x_i, y_i, t), \qquad V_i = \sum_t v(x_i, y_i, t) \quad (3)$$

Being based on a smooth optical flow, the gray levels on an optical curve change smoothly, not abruptly. Hence decomposing a GOF along these curves of regular intensity will yield less number of significant coefficients compared to normal wavelet decomposition. Such a decomposition is possible by warping the basis functions according to the optical curve. Once a GOF is formed, each frame is divided into small windows. These small windows are grouped to create sub-GOFs, as shown in Figure 1. Next, the warped wavelet coefficients are computed for each sub-GOF, and the top N significant coefficients are saved as the *Compressed Appearance Descriptors*.

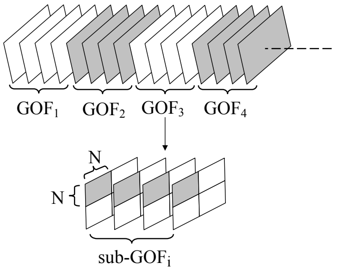


Figure 1. Creating the Group of Frames (GOF) and sub-GOFs from a sequence

The reconstruction of the data requires the knowledge of the *optical curve*. As long as the spline coefficients are known, the optical flow can be recovered, and the data can be reconstructed.

4 Applications

In this section we will demonstrate two practical applications, where the compressed spatio-temporal descriptors are used: *Video Retrieval* and *Distinguishing People in Action*. These applications essentially involve computing the similarities and dissimilarities between various types of sequences based on compressed data.

Let (M_k, A_k) be the *Compressed Motion Descriptors* and *Compressed Appearance Descriptors* for the k^{th} group of frames, where M_k is the set of all *Compressed Motion Descriptors*, and A_k is the set of *Compressed Appearance Descriptors* all sub-GOFs in GOF_k. The representation of a sequence i, consisting of L GOFs needs to be independent of the shot length so that it can be compared with shots that have different lengths. For this reason, we first compute a *Normalized Appearance Descriptor Histogram*, H^i_A, out of all A_k's $(k = 1...L)$. Then we compute $H^i_{\hat{u}}$ and $H^i_{\hat{v}}$, the *Normalized Motion Descriptor Histograms* from all M_ks. Hence, a compact representation is achieved.

4.1 Video Retrieval

Given a specific shot and a news database, i.e. an anchorman in a news program, we want to retrieve the similar anchorman shots from the database, but not the newsreels and the advertisements (Figure 2). Assuming that (M_k, A_k) are available for all files, the retrieval decision depends on the similarity of the shot representations. Therefore, given two shots i and j, this similarity is determined by the intersection (I) of their spatial and temporal descriptor histograms (4).

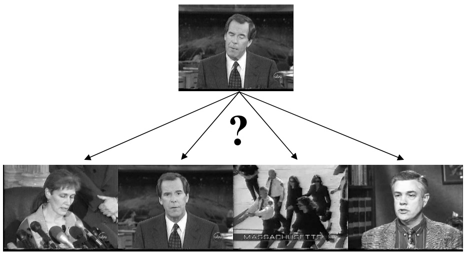


Figure 2. Given a specific shot, the aim is to select the same type of shots among many others.

$$I(i,j) = \sum_k \min(H^i(k), H^j(k)) \qquad (4)$$

Once the histograms are computed, a weighted sum of their intersections gives the desired similarity measure, as shown in (5).

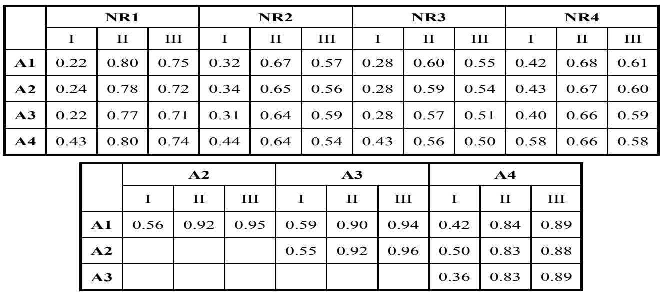
$$Sim(i,j) = w_A * I_A(i,j) + w_M * I_M(i,j) \qquad (5)$$

where $w_A + w_M = 1$ and $I_M = \frac{1}{2}(I_{M_{\hat{u}}} + I_{M_{\hat{v}}})$.

This multi-histogram technique performs much better than the simple 3D wavelet decomposition. The figure 3 shows the results of histogram intersections for various cases. In the *Case I*, the histograms are obtained from a simple 3D wavelet decomposition with no motion compensation, and the results are obviously not distinctive. However in *Case II*, although the *Normalized Appearance Descriptor Histogram* is used alone, the distinctions and similarities are clearly emphasized. The *Case III* is the inclusion of *Normalized Motion Descriptor Histogram* to Case II, improving the results more. The weighted scheme gives the luxury to change the importance of the motion and the appearance in the computations as desired, according to the shot category. For example, the anchorman shots usually involve static background and very little local motion, whereas the newsreels usually have high high global and local motion content. So when looking for anchorman shots against newsreels, it makes quite sense to increase w_M. However if the object is to find a specific anchorman, then w_A should be increased since the motion information is not distinctive anymore.

4.2 Recognizing People in Action

The spatio-temporal descriptors can be used in action recognition as well. Given a sequence of an action unit, i.e. walking and a set of sequences of unknown action units,



	NR1			NR2			NR3			NR4		
	I	II	III	I	II	III	I	II	III	I	II	III
A1	0.22	0.80	0.75	0.32	0.67	0.57	0.28	0.60	0.55	0.42	0.68	0.61
A2	0.24	0.78	0.72	0.34	0.65	0.56	0.28	0.59	0.54	0.43	0.67	0.60
A3	0.22	0.77	0.71	0.31	0.64	0.59	0.28	0.57	0.51	0.40	0.66	0.59
A4	0.43	0.80	0.74	0.44	0.64	0.54	0.43	0.56	0.50	0.58	0.66	0.58

	A2			A3			A4		
	I	II	III	I	II	III	I	II	III
A1	0.56	0.92	0.95	0.59	0.90	0.94	0.42	0.84	0.89
A2				0.55	0.92	0.96	0.50	0.83	0.88
A3							0.36	0.83	0.89

Figure 3. Results of the matching algorithm: (A: Anchorman, NR: Newsreel) The numbers show the shot similarities using the 3D wavelet decomposition with: I) no motion compensation, II) motion compensation III) motion compensation, and motion information

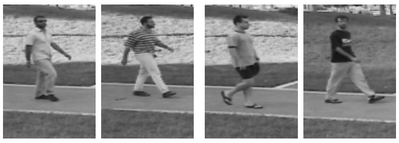
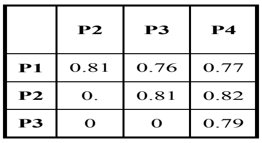


Figure 4. Sample sequences of walking action by various people

we want to find the ones that are similar to the given sequence(Figure 4). We assume that the correspondence information is available for all people in the sequences, and the image chips that contain these people is available. Once the normalized appearance and motion histograms are computed for each sequence of action unit, their intersection is computed similarly. Again, the best results are obtained for the motion compensated decomposition and the motion content as shown in figure 5.

5 Conclusion

In this paper, we present *compressed spatio-temporal descriptors* that can capture both the appearance and the motion content of a video sequence. We use a warped wavelet decomposition and spline-based optical flow estimation for this. Our approach attempts to eliminate the redundancy by finding the *optical curves* along which the image intensities are regular. Applying wavelets along these curves gives a better description of the appearance with less number of coefficients to deal with. The spline-based optical flow estimation not only can describe the motion with less number of parameters, but it also results in smoother fields of optical flow, making the *optical curves* to follow the paths along which the intensity variations are smaller.



	P2	P3	P4
P1	0.81	0.76	0.77
P2	0.	0.81	0.82
P3	0	0	0.79

Figure 5. Results of the matching algorithm for the walking sequence: (P: Person) The four sequences that contain walking action have high similarity

The video retrieval application that we present uses these small number of spatio-temporal descriptors instead of the whole data, making the retrieval process much more faster. We also demonstrate how these descriptors can be used in action recognition and recognizing specific actions. The results we get are promising and open to improvement with new measures of similarity.

References

[1] G. Mori A. A. Efros, A. C. Berg and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision*, 2003.

[2] P. Pala A. Del Bimbo and L. Tanganelli. Video retrieval based on dynamics of color flows. In *International Conference on Pattern Recognition*, 2000.

[3] D. Taubman A. Secker. Highly scalable video compression with scalable motion coding. In *IEEE International Conference on Image Processing*, 2003.

[4] M. Cascia E. Ardizzone and D. Molinelli. Motion and color based video indexing and retrieval. In *International Conference on Pattern Recognition*, 1998.

[5] A. Garrido L. Orozco-Barbosa J. Duato E. Moyano, F.J. Quiles. Efficient 3d wavelet transform decomposition for video compression. February 2001.

[6] R. Fablet and P. Bouthemy. Statistical motion-based object indexing using optic flow field. In *International Conference on Pattern Recognition*, 2000.

[7] E. Le Pennec and S. Mallat. Sparse geometric image representation with bandelets. *to appear in IEEE Transaction on Signal Processing*.

[8] J. Coughlan R. Szeliski. Spline based image registration. *Cambridge Research Laboratory, Technical Report Series*, (94/1), 1999.