

Re-covering of visual scenarios in movies by motion analysis and grouping spatio-temporal colour signatures of video shots

Jenny Benois-Pineau
IRCCyN/EPUN
jbenois@ireste.fr
rue Christian Pauc BP 50609
44306 NANTES Cedex 03 France-
LABRI UMR 5800
France

William Dupuy
IRCCyN/EPUN
wdupuy@ireste.fr
rue Christian Pauc BP 50609
44306 NANTES Cedex 03
France

Dominique Barba
IRCCyN/EPUN
dbarba@ireste.fr
rue Christian Pauc BP 50609
44306 NANTES Cedex 03
France

Abstract

The paper proposes a new, 1D signal approach to basic operations required for video structuring and video content description: motion estimation, segmenting into shots, and scene grouping. The approach is based on the projective transform called Mojette which is a discrete analog of Radon transform.

Keywords: Discrete Radon Transform, Colour similarity measure, Chain clustering.

1 Introduction

The storage and efficient retrieval of video documents in digital form in digital libraries suppose that a preliminary structuring and indexing of them by content is realized. The new coming standard MPEG7[7] requires structuring of video documents into logical video segments, such as shots, scenes, objects and indexing them by characterizing their motion, colour, texture. This task can be qualified as re-covering of a visual scenario of a video document in the sense that it allows an extraction of its basic elements: shots and scenes. Based on analysis of methods of video partitioning into shots, scene grouping [3], it can be stated that a set of various techniques has to be proposed to efficiently accomplish these three tasks. Motion and colour are necessary descriptors of video content which allow for its segmentation and even semantic

characterisation knowing content nature (sport, news, documentary, romance etc.) The difficulty to use these features consists in a strong computational cost for estimation of global motion and computation of color descriptors of 2D frames. In our previous work [5] we proposed to realise motion analysis in 1D projective transform domain, called Mojette [9], which is a discrete analog of Radon transform used in tomography. Measures of similarity of frames based on estimated motion in 1D Mojette domain allow for shot change detection. In this paper we further develop motion estimation methods, introduce colour similarity measures in Mojette transform domain and show how elements of a visual scenario such as shots and scenes can be extracted from raw video in 1D signal domain. The paper is organised as follows. In Section 2 motion estimation and shot change detection in Mojette transform domain are developed. In section 3, scene grouping method in Mojette transform domain is described, section 4 describes experimental results and outlines perspectives of this work.

2 Motion estimation in Mojette transform domain

Affine models of apparent motion in 2D image plane were proved to be interesting for the characterisation of camera work in video. Many authors limit themselves to a 3 parameter motion model [8] with translational and zoom factors as it allows for a characterisation of most frequent situations in video. This 2D motion model can be calculated by

estimation of 1D motion model in Mojette transform domain.

2.1 Mojette Transform

Mojette transform introduced in [9] is a discrete version of the Radon transform used in tomography.

Supposing a Dirak model of image pixels, it is given by a discrete signal

$$M_{p,q}[J](m) = \sum_{k,l} I(k,l)\delta(m+q.k-p.l) \quad (1)$$

where $I(k,l)$ is the image intensity value at pixel coordinate (k,l) , δ is a Kronecker symbol and m is the Mojette index. Each element $M(m)$ of the transform, also called "bin", is the result of the summation of a set of intensity pixels along the line defined by:

$$m - q.k + p.l = 0 \quad (2),$$

where p and q are mutually prime integers.

2.2 Motion estimation in Mojette transform domain.

Assuming a 3 parameter affine 2D motion model in image plane, the elementary displacement vector $(dx,dy)^T$ at each pixel position $(x,y)^T$ is expressed as

$$\begin{cases} dx = t_x + f(x - x_g) \\ dy = t_y + f(y - y_g) \end{cases} \quad (3)$$

here $(t_x, t_y, f)^T$ are the parameters of the model, respectively the horizontal translation, the vertical translation, and the isotropic zoom, and $(x_g, y_g)^T$ is the reference point of model, generally the center of the image. In [5] we showed that this 2D motion model corresponds to a 1D motion model, where the elementary displacement in Mojette transform domain can be expressed as

$$dm = t_m + f(m - m_g) \quad (4)$$

where $t_m = -q.t_x + p.t_y$ is the transformed translation vector, and $m_g = -q.x_g + p.y_g$ is the transformed reference point. The zoom factor remains the same as in 2D case.

To estimate 3 parameter affine model at least two non-co-linear directions of Mojette projection should be used. Then if we denote M_j ($j=1,2$) two projections with direction (p_j, q_j) and tm_j their corresponding 1D translations, then the 2D translation parameters can be recovered from the linear system.

$$\begin{cases} t_{m1} = -q_1.t_x + p_1.t_y \\ t_{m2} = -q_2.t_x + p_2.t_y \end{cases} \quad (5)$$

Taking a set of $n > 2$ directions, the model equation can be written as

$$\begin{pmatrix} tm_1 / \sqrt{p_1^2 + q_1^2} \\ \dots \\ tm_n / \sqrt{p_n^2 + q_n^2} \end{pmatrix} = \begin{pmatrix} -q_1 & p_1 \\ \sqrt{p_1^2 + q_1^2} & \sqrt{p_1^2 + q_1^2} \\ \dots & \dots \\ -q_n & p_n \\ \sqrt{p_n^2 + q_n^2} & \sqrt{p_n^2 + q_n^2} \end{pmatrix} \begin{matrix} / \\ \\ \\ / \\ \end{matrix} \times \begin{pmatrix} t_x \\ t_y \end{pmatrix} \quad (6)$$

Then the parameter vector $t = (t_x, t_y)^T$ can be estimated by a classical Least Square estimator. The zoom factor is computed as a mean value of zoom factors f_i estimated for each direction of projection.

To estimate the parameter vector $(tm, f)^T$ (4) in Mojette transform domain, we propose a robust estimator based on local correlation of two Mojette projections of the same directions of two successive frames in a video sequence $M_{p,q}^t, M_{p,q}^{t+1}$:

$$\Psi(dm) = \sum_m \gamma(\rho_j(m+dm)) \quad (7)$$

Here dm is an elementary displacement (3) in Mojette transform domain, ρ_j is a correlation coefficient computed in a window around j -th bin of $M_{p,q}^t$, γ is a robust function derived from Tuckey estimator [5]

$$\gamma(x) = \begin{cases} 1 - \frac{(x-1)^6}{(1-C)^6} + 3 \frac{(x-1)^4}{(1-C)^4} - 3 \frac{(x-1)^2}{(1-C)^2} & \text{if } C < x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

with C a predefined constant.

2.3 Shot change detection

The correlation of successive frames after motion compensation in Mojette domain shows the similarity of video content. The lack of correlation depicts its changes. Thus shot changes can be detected based on this measure. In our previous work [5] we proposed to detect shot changes in video as strong decrease of the mean value $\bar{\rho}$ of global correlation coefficient between Mojette projections of successive frames. The detection of jumps in mean $\bar{\rho}(t)$ is realized by an on-line statistical Hinkley test described in [4]. In the actual work we tested the detection of jumps in the mean value of robust functional (7). The behaviour is

similar and allows for detection of shot change and transition effect identification (cut/progressive change) even in the case of complex effects. An example of detection in a publicity clip @SFRS is given in Figure 1. Here, the detection rule is as follows The first upward jump excluded, a pair of successive downward and upward jumps correspond to a shot change, only the first of two successive jumps of the same sign is considered.

3 Scene grouping based on colour signatures of shots.

While shot change detection techniques are based on a break of similarity of time continuous video content, the scene grouping methods aim at clustering shots with similar content into scenes. Continuity of the colour in temporarily adjacent shots is a characteristic feature of such a large class of video content as documentaries and also "author movies", where the colour is a strong stylistic tool. In [1] a method for measuring of shots similarity based on Vector Quantisation was proposed using LVQ method. Namely for each shot its own VQ dictionary is designed, where each vector is a block in each video frame of the shot. Then the errors of representing shots with dictionaries are computed. Thus for a shot S_j the error $D_i(S_j)$ is the error of representing of the shot S_j with the dictionary designed for the shot S_i . The use of Mojette transform in this context is very interesting. Preserving spatial distribution of colour in image plane, it allows a reduction of the order of computational complexity. Thus for a block-vector of dimension of n^2 and for the projection direction (p,q) the dimension of the vector representing the projection of the block is $(n-1) * (|p| + |q|)$ as it can be derived from the result in [9]. The colour projection vector can be constructed in YUV and HSV colour systems and represents a concatenation of three vectors each for one system component. Based on coding distortion, a similarity measure for two video shots proposed in [1] is

$$e(S_i, S_j) = |D_j(S_i) - D_i(S_i)| + |D_i(S_j) - D_j(S_j)| \quad (8)$$

$$\text{where } D_i(S_j) = \sum_{k=1}^{N_B} d(\text{Block}_{j,k}, \text{Dico}_i)$$

with

$$d(\text{Bloc}_{j,k}, \text{Dico}_i) = \sum_{l=1}^{\text{VectorDimension}} (x_l - V(x_l))^2$$

Here $V(x)$ is the vector in the dictionary which

codes the original vector x . We call spatio-temporal colour signature of a video shot S_i its dictionary Dico_i in colour Mojette transform domain. In our work we used a SplitLBG [6] method which is a generalized K-means clustering to construct the dictionary for each shot.

The method of grouping shots in video scenes we propose uses spatio-temporal signatures of shots measuring shots similarity (8). The method is based on two main assumptions : (i) when the measure $e(S_i, S_j)$ is small, the shots are similar in colour space and (ii) shots constituting the same scene are closed in time. Thus the grouping method uses both colour similarity of shots and temporal distance between them to group shots into clusters corresponding to scenes. The idea to introduce temporal constraint into shot grouping algorithm was firstly proposed in [10]. Here a temporal distance between shots is defined as

$$d_{temp}(S_i, S_j) = \min(|n_b(S_i) - n_e(S_j)|, |n_b(S_j) - n_e(S_i)|),$$

where n_b is the number of initial frame in the shot and n_e is the number of the final frame. In [10] a time window is introduced and only shots with a time distance inside the time window can be grouped into the same scene. In our work we did not use such an abrupt constraint and propose to introduce a weighted spatio-temporal similarity measure or weighted distance:

$$d(S_i, S_j) = d_{temp} * e(S_i, S_j) \quad (9)$$

Based on this weighted distance, shots can be grouped into clusters which express both colour and temporal proximity of shots. Thus they correspond to scenes in movies according to our model of content. As for clustering method to be applied, it is clear that methods with fixed number of clusters are not applicable, as the aim here is to find spatio-temporal borders of an unknown number of scenes in a movie. This is why we use a "chain clustering" method which we studied in detail in [2]. The principle of this method consists in iterative ordering of all objects in the data space into an ordered set called "chain" c . At i -th iteration chain c_i is constructed by adding to the chain c_{i-1} the object o^* from initial object set such that the distance $d(o^*, c_{i-1})$ is minimal for all objects remaining in initial data set and not yet joined to the chain. We recall that the distance between an object and a set (chain) is the minimal distance between the object and all elements of the set. In [2] we showed how this method can be implemented in $O(N^2)$ operations. After the chain c is constructed, the splitting of it

into clusters is done by thresholding the distance $d(o^*, c_{i-1})$ defined on chain elements or by detecting its maxima and cutting in maxima locations. The second manner is more interesting as it allows an “adaptive” clustering. In our work we propose firstly to filter the distance function by low-pass filtering and then to cut the chain. Figure 2. shows the filtered distance function for a chain of 100 shots.

4 Results and Perspectives

Proposed methods of extraction of elements of visual scenarios in video documents were tested on a data set containing documentaries and movies: MPEG7 dataset, Ina © data set “Avengers”, “Histoire d’eau”, documentaries SFRS ©. Some illustrations are given on Figures 1, 2, 3.

As far as motion estimation is concerned, the 30% increase of Peak Signal to noise ratio after motion compensation by estimated global model (see section 2) shows that the method in 1D signal domain gives comparable results as classical 2D methods.

The shot change detection compared to the ground truth gives recall of 98% and precision of 80% approximately. It strongly depends on textural content of video frames.

As far as grouping method is concerned, the illustration of its result is given in Figure 3. The recall here is of 100% and precision is of 73%. It can be explained by the use of HCV colour system and the sensitivity of proposed similarity measure to changes of lightening conditions.

In conclusion we will state that fundamental operation of video content structuring, such as shot change detection and scene grouping can be successfully done in 1D projective transform domain, which preserves spatial and colour structure of video frames. These studies are far from being exhausted. This approach opens a fascinating perspective from the computational point of view reducing combinatorial complexity of basic video analysis operations.

References

[1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L. Rossi, “Describing multimedia documents in natural and semantic-driven ordered hierarchies”. In Proc; ICCASP’2000, Istanbul, Turkey, 6-9 June 2000.

[2] J. Benois-Pineau, A. Khrennikov, N. Kotovitch. « Image segmentation in compressed domain by clustering methods with euclidean and p-adic metrics », submitted to IEEE PAMI.

[3] Del Bimbo A., Visual Information Retrieval, Morgan Kaufman Publishers, Inc, San Francisco, California, 1999

[4] P. Bouthemy, M. Gelgon, F. Ganausia “A unified approach to shot change detection and camera motion characterization”, IEEE Circuits and Systems for Video Technology, october 99, vol 9, n°7, pp. 1030-1044

[5] F. Coudert, J. Benois-Pineau, and D. Barba, “Dominant motion estimation and video partitioning with a 1D signal approach”, Proc of SPIE Intl. Conf. on Voice, Video and Data Communication, SPIE 3127, Boston, November 1998.

[6] R.M.Gray, “Vector Quantization”, IEEE ASSP magazine, pp4-28, April 1984

[7] ISO/IEC JTC 1/SC 29/WG 11/M6156, MPEG-7 Multimedia Description Schemes WD (Version 3.1) July 2000, Beijing

[8] P. Joly, H.-K. Kim, “Efficient automatic analysis of camera work and micro-segmentation of video using spatio-temporal images”, Signal Processing : Image Communication , 8(1996), pp. 295-307

[9] N. Normand, J.-P. Guedon, “La transformée Mojette: une représentation redondante pour l’image”, C. R. Acad. Sci. Paris, t. 326, Série I, pp. 123-126, 1998

[10] M.M. Yeung, B-L Yeo, « Time-constrained Clustering for Segmentation of Video into Story Units », Proceedings of ICPR Vol.3, pp. 375-380, August 1996.

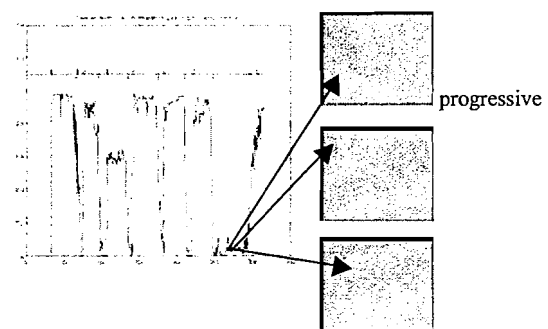


Figure 1: Shot change detection with “Cut” and progressive effects. “Logo” clip SFRS©.

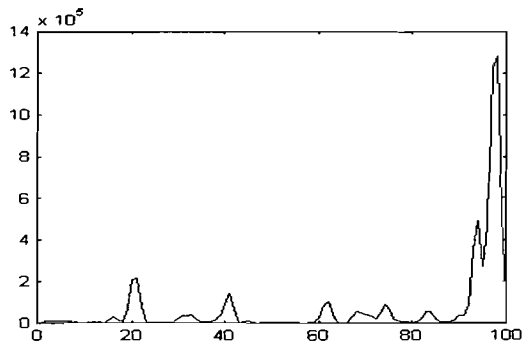


Figure 2. Spatio-temporal similarity measure or weighted distance for a 100-shots chain. "Avengers" clip INA®.

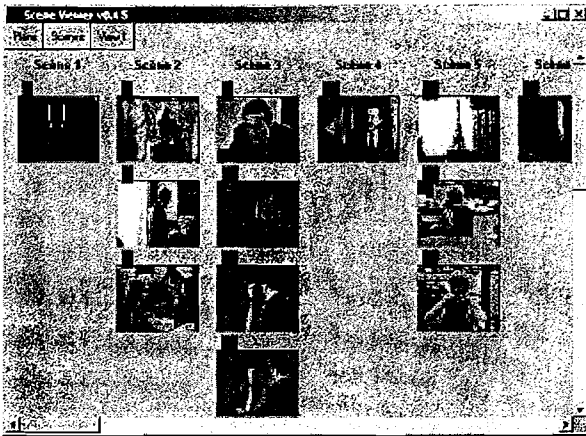


Figure 3: Fragment of the scene spatio-temporal structure. "Avengers" clip INA®.