ALIGNMENT OF NEARLY-REPETITIVE CONTENTS IN A VIDEO STREAM WITH MANIFOLD EMBEDDING

Manal Al Ghamdi *Yoshihiko Gotoh*

Department of Computer Science, University of Sheffield, United Kingdom
Email: {m.alghamdi,y.gotoh}@dcs.shef.ac.uk

ABSTRACT

This paper presents an approach to identifying nearly repetitive contents in a stream of video where prior information such as the number, the length and contents of repetitions are not known. The approach is novel in that it does not require a template for searching or learning repeated contents. Instead it analyses a video by characterising the spatial and temporal information embedded in a frame sequence. A video is represented with its spatio-temporal features, which are analysed in the embedded manifold to reconstruct the underlying structure so that repeated contents can be reorganised. The approach is evaluated using rushes videos, where numerous repetitions are found. The experiments show that overall performance is improved using the extension of manifold learning with the spatio-temporal representation.

Index Terms— spatio-temporal representation, manifold, synchronisation, inter-similarity, rushes video.

1. INTRODUCTION

Repetitive contents in multimedia are frequently found in a combination of textual, visual and audio (speech) information. A quick search for any multimedia materials using conventional search engines often results in multiple items with similar, or even identical, contents in the highest rank. In news broadcasts, for example, we frequently see nearly-repeated video footage although the presentation may vary with, *e.g.*, camera settings and appearance of objects, reflecting their production processes and policies. Repetitive contents are not copies, but there exist some differences, thus making their management a difficult problem.

Rushes video is one example of nearly-repetitive sequences, whereby the original material is transformed into nearly, but not exactly, identical contents. It contains repetitive contents from multiple retakes of the same scene, caused by, *e.g.*, actors' mistakes or technical failures [1]. Nearly-repetitive contents in the rushes video may not be identical, sometimes causing inconsistency between retakes [2]. Occasionally some parts of the original sequence are dropped or extra information may be added at various places, resulting in retakes of the same scene with unequal lengths.

In this paper we present an approach to align nearly-repetitive contents in rushes video. The majority of the previous works employed techniques such as template matching, temporal windowing, segmentations, camera calibration analysis or object tracking. Instead, we define a spatio-temporal inter-similarity between repeated video contents by extending the spatial Isomap to spatio-temporal graph-based manifold embedding (or STG-Isomap) that captures the similarities between repetitive sequences. Firstly spatio-temporal features are defined for a high-level semantic representation of complex scenes in a video sequence. Interest points that have significant local variations in both space and time are extracted and encoded using fewer codebook basis in the high-dimensional feature space. We used the spatio-temporal extension of locality-constrained linear coding (ST-LLC) that is able to detect features using spatio-temporal scale-invariant feature transform (ST-SIFT) [3]. At each time instance (a video frame, practically) visual features, defined by the ST-LLC codes, are modelled to form a temporal coherence to adjacent frames. Secondly, the similarity is measured by constructing a shortest-path graph with k-nearest neighbour (kNN) in both the spatial and the temporal domains. We introduce an extension of Isomap aiming to identify the underlying structure in repetitive video sequences, semantically modelled by ST-LLC codes. The structure of heterogeneous data is reconstructed, presenting clusters of repetitive scenes.

2. RELATED WORK

Manifold learning is a class of non-linear dimensionality reduction technique that transfers data from a high-dimensional space to a suitable output space with reduced dimensionality [4]. Non-linear manifold learning does not assume the linearity of the input space, thus providing a better chance of dealing with input data with complex embedding in the high-dimensional space. The space with reduced dimensions should reflect the intrinsic dimensionality of the data, that is, the least number of parameters that capture the data features.

Manal Al Ghamdi would like to thank Umm Al-Qura University, Makkah, Saudi Arabia for funding this work as part of her PhD scholarship programme.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

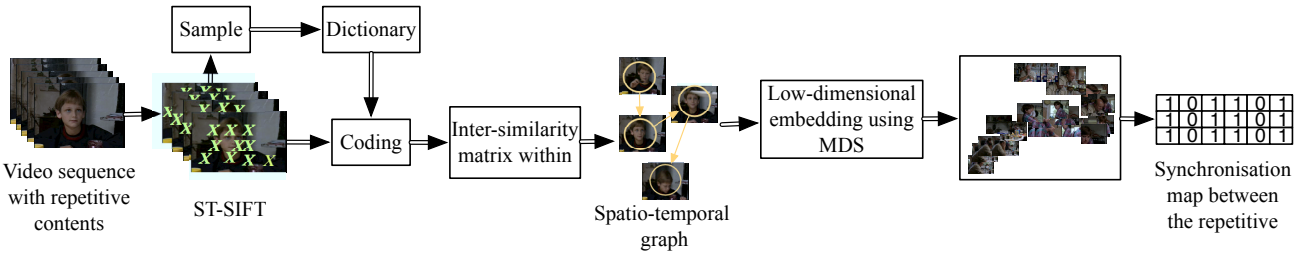


Fig. 1. Processing steps for nearly-repetitive contents alignment.

There are many manifold learning algorithms adopted in the video processing field, among which isometric feature mapping (Isomap) is one of the widely used techniques. It is a graph-based scheme that compounds the highlighted features of the principal component analysis (PCA) and multidimensional scaling (MDS) [5]. It overcomes dimensionality reduction problems by considering the geodesic distance, which defines the shortest path or curve that connects two points in a connected set.

Various problems, such as sequences alignment, have been addressed using manifold techniques. Because they involve time series data, one potential direction would be to exploit the temporal information in learning the reduced-dimensionality space. As far as we are aware, in the manifold learning literature, only Jenkins and Matarič [6] took temporal coherency explicitly into account. Their algorithm extended the spatial Isomap [5] by assigning similar low-dimensional weights to temporally adjacent samples extracted using a windowing technique. They grouped these samples so that temporally adjacent groups would have similar low-dimensional coordinates. They did not model dynamics and their performance depends on the window size, where smaller windows produced better results.

3. ALIGNMENT OF REPETITIVE CONTENTS

Our work defines the relation between nearly-repetitive contents in a video stream (*i.e.*, multiple retakes of the same scene) in the low-dimensional space. The approach consists of two stages. First, content of the video stream is described in the high-dimensional space using the invariant interest points and coding schemes (Section 3.2). To define spatio-temporal codes that represent video frames we apply ST-LLC which considers the locality of the manifold structure in the input space [3]. Second, a manifold is computed in order to map the high-dimensional representation to the embedded space (Section 3.3). At this stage the inter-similarity is computed between the multiple retakes using the spatio-temporal kNN graph. We extended the spatial Isomap [5] to the spatio-temporal domain to generate the intrinsic coordinates for each manifold. Generated coordinates are chronologically ordered by the spatio-temporal similarity and integrated to a graph for sequences alignment. The entire process of the approach is illustrated in Figure 1.

3.1. Notation

We use the following notations in the remaining sections of this paper. Let $X = \{x_1, \ldots, x_N\} \in \mathbb{R}^{D \times N}$ denote a video sequence containing repetitive contents with N frames and D dimensions, and x_i represents a frame in X. $F = \{fx_1, \ldots, fx_N\} \in \mathbb{R}^{Q \times N}$ is the ST-SIFT features with N entries and Q dimensions, where $fx_i = \{f_1, \ldots, f_M\} \in \mathbb{R}^{Q \times M}$ represents a set of M interest points for frame x_i. Let $S = \{sx_1, \ldots, sx_N\} \in \mathbb{R}^{U \times N}$ be the spatio-temporal codes that represent the video frames with N codes and U dimensions, and $sx_i = \{s_1, \ldots, s_M\} \in \mathbb{R}^{R \times M}$ represents a set of M codes for the frame x_i.

3.2. Video Representation

The first step is to apply the ST-SIFT algorithm that identifies spatially and temporally invariant interest points in a given video stream. These points contain the amount of information sufficient to represent the video contents. To achieve the invariance in both space and time, spatio-temporal Gaussian and difference of Gaussian (DoG) pyramids are calculated. The set of points F shared between three spatial and temporal planes (xy, xt and yt) at each scale in the DoG are chosen as interest points. The second step is to derive the spatio-temporal codes S for the video stream X given the ST-SIFT feature matrix F. For each frame x_i, the algorithm works by firstly constructing a spatio-temporal graph between its descriptor set fx_i and a codebook B, computing the shortest path, performing a kNN search, and finally solving the following constrained least-square fitting problem:

$$\min_{sx_i} \sum_{j=1}^{M} \|f_j - Bs_j\|^2 + \lambda \|d_j \odot s_j\|^2 \quad st. \quad 1^\top s_j = 1, \ \forall j$$

where $\odot$ is the element-wise multiplication, λ is a sparsity regularisation term and d_j is the locality parameter that represents every basis vector with different freedom based on

its shortest path to the spatio-temporal descriptor. '$1^\top s_j = 1,\ \forall j$' is the shift-invariant requirements for the LLC code. The final step uses the multi-scale max pooling [7], where the set of codes computed for each frame are grouped together to create the corresponding pooled representation S.

3.3. Manifold Learning

Given a spatio-temporal coding matrix S for a video sequence X, the synchronisation map is estimated between the multiple video retakes. First the similarity matrix δ is calculated between the LLC codes using the Euclidean distance. The value of δ_{ij} defines the distance between the LLC codes sx_i and sx_j for two frames x_i and x_j $(i, j = 1, \ldots, N)$. Then, for each frame instance x_i represented by a code sx_i:

1. L frames whose distance is the closest to x_i are connected. They are referred to as spatial neighbours (sn):

$$sn_{x_i} = \left\{ sx_{j1}, \ldots, sx_{jL} \mid \underset{j}{\operatorname{argmin}}^{L}(\delta_{ij}) \right\} \quad (1)$$

   where $\underset{j}{\operatorname{argmin}}^{L}$ implies node indexes j with L smallest distances.

2. Another L frames, chronologically ordered around x_i, are set as temporal neighbours (tn):

$$tn_{x_i} = \left\{ sx_{i-\frac{L}{2}}, \ldots, sx_{i-1}, sx_{i+1}, \ldots, sx_{i+\frac{L}{2}} \right\} \quad (2)$$

3. To optimise the set of temporal neighbours, tn_{sn} is selected from temporal neighbours of spatial neighbours:

$$tn_{sn_{x_i}} = \left\{ tn_{s_{j1}}, \ldots, tn_{s_{jK}} \right\} \cap tn_{s_i} \quad (3)$$

4. Spatial and temporal neighbours are integrated, producing spatio-temporal neighbours (stn) for frame x_i:

$$stn_{x_i} = sn_{x_i} \cup tn_{sn_{x_i}} \quad (4)$$

   The above formulation of stn_{x_i} effectively selects x_i's temporal neighbours that are similar, with a good chance, to its spatial neighbours. This means that, suppose x_i is an isolated frame and totally different from the temporal neighbours, only the spatial neighbours will be taken into consideration.

The inter-similarity matrix is constructed by recalculating the shortest path between the nodes in graph δ, forming a new embedded correlation δ_γ. The manifold embedding is then modelled as a transformation T of the high-dimensional data in terms of similarity δ_γ into a new embedded configuration E in the low-dimensional space:

$$T : \delta_\gamma \ \rightarrow E \quad (5)$$

The function T is the eigen decomposition of the inter-similarity matrix that minimises the following loss function:

$$\begin{aligned} L_{projection} &= \|X - T(X)\| = \|X - T(\delta_\gamma)\| \\ &= \|X - (Q \wedge Q^T)\| = \|X - (Q_+ \wedge_+^{\frac{1}{2}})\| \end{aligned} \quad (6)$$

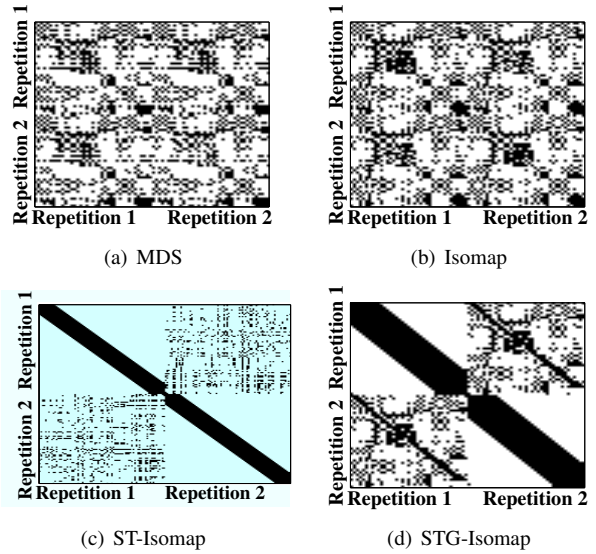
where Q and $\wedge$ are the eigenvectors and the eigenvalues of δ_γ, $\wedge_+^{\frac{1}{2}}$ contains the e largest eigenvalues in $\wedge$ along the diagonal and Q_+ is the square root of e columns in Q. The new coordinates for each frame instance in the embedded space are selected from the e largest eigenvalues of matrix $Q_+ \wedge_+^{\frac{1}{2}}$.

4. EXPERIMENTS

The approach was evaluated using MPEG-1 videos from the NIST TRECVID 2008 BBC rushes video summarisation task [1]. Three video sequences identified as *MS206370*, *MRS150072* and *MRS044499* were selected for evaluation, with the approximate length of two minutes, the frame rate of 25 fps (frames per second) and the frame size of 288×352 pixels. Video representation was created as follows: Firstly we used spatio-temporal SIFT combined with LLC [3] as a local features detector. Using publicly available code by Scovanner *et al.* [8], spatio-temporal regions around the interest points were detected and described by the 3D-HOG (histogram of Gaussian). For each interest point the descriptor length was 640-dimensional and was determined by the number of bins to represent the orientation angles in the sub-histograms. To create a pooled representation in the SPM (spatial pyramid matching) step, the ST-LLC was computed for each spatio-temporal sub-region and pooled together using multi-scale max pooling. We used 4×4, 2×2 and 1×1 sub-regions. The pooled features were then concatenated and normalised using the ℓ^2-norm. For dictionary generation, descriptors for interest points were clustered to a pre-specified number of visual words. The similarity matrix was computed using the Euclidean distance and the spatio-temporal kNN graph was constructing with $k = 20$. Lastly the classical MDS was applied for dimensionality reduction.

4.1. Evaluation Schema

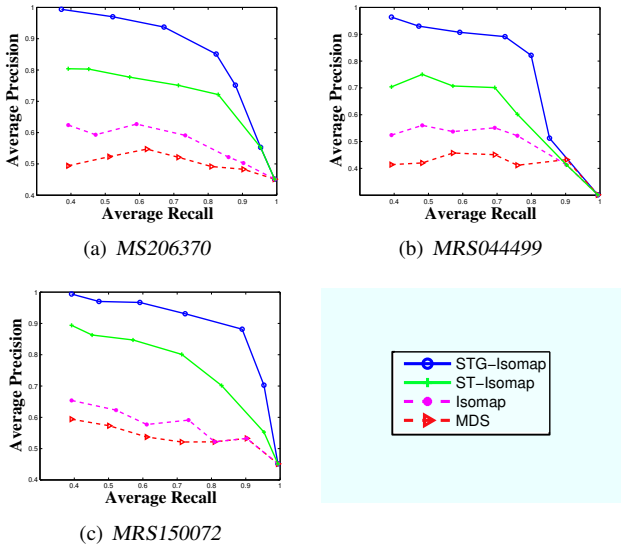
Each scene in the rushes videos is a sequence of actions defined by actors' dialogue. A scene was used as a unit for evaluation. The purpose of the experiment was to align and group the multiple and similar retakes of the same scene. A description of actions for each scene was provided by NIST [1]. The groundtruth was created using three human judges. To measure the performance of retakes similarity, the average precision and recall were calculated and compared with three other approaches including MDS, the conventional Isomap and the ST-Isomap [6]. Individual performances were evaluated using a kNN graph with the varying value of k.



(a) MDS (b) Isomap

(c) ST-Isomap (d) STG-Isomap

Fig. 2. Synchronisation binary map to illustrate the relation between two repetitive sequences in *MRS044499*.



(a) *MS206370* (b) *MRS044499*

(c) *MRS150072*

Fig. 3. Average precision and recall for three rushes videos. The STG-Isomap approach is compared with the MDS, the conventional Isomap and the ST-Isomap.

4.2. Results and Analysis

This experiment aimed at reconstructing video sequences to uncover their repetitive contents. Four manifold learning techniques were applied on the spatio-temporal representation. It was hoped that retakes from the same scene would be mapped close to each other in the manifold, resulting in clusters of repetitive contents. Pairs of closest frames from the different retakes would form an alignment. Figure 2 illustrates the synchronisation map computed between two nearly-repetitive sequences in rushes video *MRS044499*. The figure shows that the STG-Isomap was able to build cleaner clusters of repetitive contents, with which most frames from the same scene were re-positioned and closely placed in the embedded space. It developed the spatio-temporal relation during neighbourhood graph construction. As a comparison, only the temporal relation was identified by the ST-Isomap while the spatial relation was identified by the Isomap and the MDS.

For detection of similar and repetitive scenes, precision and recall scores are presented in Figure 3. The figure indicates that the spatio-temporal approach outperformed other methods with all three videos. The inter-similarity was considered as a multi-region frame-by-frame comparison, calculating the similarity of composition between two frames using the spatio-temporal codes. This is because, in a sense, individual objects' characteristics were coherently transitioned from one frame to another. The difference was not very large between the MDS and Isomap for embedding repetitive manifolds, indicating that these methods could not capture spatio-temporal features very well. As an additional note, the STG-Isomap chose the value for L before the graph was constructed. For the Isomap and ST-Isomap, on the other hand, L was selected during the graph construction, and they could change the value when calculating the shortest paths. The linear embedding technique, MDS, did not have a step for graph construction, but the fixed neighbourhood size was applied. It was not able to learn the repetitive multi-class sequences very well.

One may notice that Figure 3(c) for video *MRS150072* presented the best results. It was caused by variations in scene setting such as appearance of dominant colour patterns and moves of video shooting location in the scene. Large variations contributed particularly well for identification of repetitive contents. On the other hand, Figure 3(b) for video *MRS044499* resulted in lower performance across various approaches. It was an indoor scene with little moves by actors and, despite the use of different camera angles in the retakes, there were no dramatic changes with colour combination of the foreground (*i.e.*, actors themselves) and background.

5. CONCLUSIONS

This paper presented an approach to aligning nearly-repetitive contents in a video stream using manifold embedding. It utilised ST-LLC to densely extract and encode salient feature points from a 3D signal, capturing the intra-similarity within the video sequence. A spatio-temporal graph was derived as a step for manifold learning that defined the inter-similarity across two sequences. Experimental results using rushes video showed that the presented approach performed better than the conventional manifold embedding techniques. The contribution of this study may be extended to other temporal applications such as video content similarity and video information retrieval.

6. REFERENCES

[1] P. Over, A. F. Smeaton, and G. Awad, "The TRECVID 2008 BBC rushes summarization evaluation," in *Proceedings of the TRECVID Video Summarization Workshop*, 2008.

[2] A. Joly, O. Buisson, and C. Frelicot, "Content-based copy retrieval using distortion-based probabilistic similarity search," *IEEE Transactions on Multimedia*, vol. 9, no. 2, 2007.

[3] M. Al Ghamdi, N. Al Harbi, and Y. Gotoh, "Spatio-temporal video representation with locality-constrained linear coding," in *Computer Vision*, 2012.

[4] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: a comparative review," Tech. Rep., Online preprint, 2008.

[5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, 2000.

[6] O. C. Jenkins and M. J. Matarić, "A spatio-temporal extension to isomap nonlinear dimension reduction," in *Proceedings of the ICML*, 2004.

[7] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proceedings of the IEEE CVPR*, 2005.

[8] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proceedings of ACM Multimedia*, 2007.