NEARLY-REPETITIVE VIDEO SYNCHRONISATION USING NONLINEAR MANIFOLD EMBEDDING

Siripinyo Chantamunee Y Yoshihiko Gotoh

University of Sheffield, Department of Computer Science
211 Portobello Street, Sheffield S1 4DP, United Kingdom
Email: {s.chantamunee, y.gotoh}@dcs.shef.ac.uk

ABSTRACT

This paper introduces a new method for synchronising nearly repetitive video sequences using manifold embedding. The approach recovers the optimal correspondences between two frame sequences in the embedded space. The synchronised map is estimated based on the spatio-temporal extension of Isomap. It is evaluated using the TRECVID rushes video collection. The accuracy for repetitive shot synchronisation is scored using precision and recall of all generated edges on synchronised graph. We show that the approach significantly improves the embedding for repetitive manifolds over the well investigated methods such as multidimensional scaling, locally linear embedding, and the conventional Isomap.

Index Terms— Synchronization, Video signal processing, Multidimensional sequences, Visualization.

1. INTRODUCTION

The rapid growth of multimedia technology in recent years has caused an exponential increase of multimedia digital data, including broadcast news, TV programmes, movies, dramas, meetings, and sport video. The pervasive availability of audio-visual copy equipments gives rise to potential problems for content management. A nearly identical material is repeatedly produced, aiming at transformation of an original material into an another form serving different purposes. In video industries, high-level postprocessing is performed before distribution: insertion of logos, picture-in-picture, fingerprinting, rescaling, and cropping [1].

Rushes video — sometimes referred to as pre-production video — is a raw material that will be used to create multimedia products such as drama and television programmes [2]. It is a collection of retakes of the same scene. Low-level transformation on the same contents can be performed during production using different camera angles and view settings. Occasionally part of contents are dropped, or extra information may be randomly added. A human factor adds further unreliability (e.g., actors often make mistakes, or perform along different lines at each retake) causing different lengths of the same scene.

Rushes video consists of nearly replicated frame sequences instead of verbatim copy [3]. A replica may not be identical to the original, therefore causing a certain amount of inconsistency between retakes. Transformation can occur at various levels, from low to high, causing problems in identification of duplicated video contents. The ambiguity of repeated spatial information during the temporal changes attracts further challenges. Technologies for modelling video sequences using spatial and temporal information may bring significant advantages for video content management task.

In this paper we present a novel approach to video synchronisation problem for nearly repetitive sequences in rushes video. Unlike other works (e.g., [4][5][6]) the approach does not require frame-to-frame feature matching, camera calibration analysis or scene object tracking. Instead, it extracts the spatio-temporal correlation within and across nearly replicated video sequences and generates a series of intrinsic coordinates on the embedded space. As a consequence identification of corresponding frames across repetitions becomes a simpler task. It is the advantage over the traditional techniques for manifold embedding such as Isomap (isometric feature mapping), LLE (locally linear embedding) and Laplacian eigenmaps [7] because they do not address the problem of integrating repeated manifolds in terms of spatial and temporal relationship; they find low-level coordinates only when geometry representation of a manifold is preserved.

We aim to develop an extension of graph-based manifold embedding – Isomap – to capture spatio-temporal correlations between repetitive sequences in the embedded space. Repetitive sequences are modelled with the GMM (Gaussian mixture model) and correlation between frames are estimated using the MLE (maximum likelihood estimation). Synchronisation will be achieved from the shortest paths generated during the embedding. Experimental results indicate that the spatio-temporal extension of Isomap significantly improves the embedding for repetitive manifolds over the conventional approaches.

2. RELATED WORKS

Manifold learning is the process for estimating the underlying structure within high-dimensional data. A number of intrinsic coordinates on the low-dimensional space are generated while preserving their original geometry. In recent years there exist growing interests in nonlinear manifold learning to deal with complex high-dimensional data and applications. Various techniques has been proposed: MDS (multidimensional scaling), Isomap, kernel PCA (principal component analysis), LLE, diffusion maps, Hessian LLE, etc [7]. MDS is the traditional technique for approximating the configuration of embedded coordinates from the Euclidean pairwise distance matrix. Isomap is a graph-based embedded method whereby a graph is created by computing the k-nearest neighbours. The geodesic distance is then calculated in order to generate a new distance graph. However only a spatial geometry structure is preserved. LLE is similar to Isomap, but it estimates the weights of local neighbourhoods for every point in the graph instead of calculating geodesic distance. These techniques have been extensively studied in several applications such as mobile localisation [8][9], dimension reduction [7], visualisation and alignment for manifold information [10]. In this paper we in-
vestigate the problem of synchronising nearly repetitive sequences on manifold embedding.

3. MANIFOLD LEARNING FOR REPETITIVE SEQUENCES

The problem of sequence synchronisation is framed as finding correspondences between two nearly repetitive manifolds that are coherently embedded on a low-dimensional space. The approach consists of the following two stages. Firstly, in a high-dimensional space, spatio-temporal correlation is derived from frame sequences to represent similarity across frame instances in terms of spatial and temporal coherence (Section 3.1). Specifically, two types of correlation — intra and inter-sequence correlation — are computed for further use during the sequence synchronisation stage. Secondly, for each sequence, a video scene manifold is learnt in order to map a high-dimensional video sequence into an embedded dimensional space. A series of intrinsic coordinates for each manifold is generated using the spatio-temporal extension of Isomap (Section 3.2). Manifold instances are chronologically ordered and coherent by context of spatial and temporal similarity. Two manifolds are then iteratively integrated to create a graph for sequence synchronisation. The illustration for the process of our approach is shown in Figure 1.

\[
L_{projection} = ||\delta - T(\{\delta_x, \delta_y, \delta_{xy}\})||
= ||\delta - T(\delta_{\gamma})||
= ||\delta - (Q\wedge QT)||
= ||\delta - (Q^+\wedge Q^T)||
= ||\delta - (Q^+\wedge \frac{1}{2})||. \tag{4}
\]

Fig. 1: The processing steps for Video Synchronisation using Manifold Learning.

Mathematical development for the proposed approach is outlined as follows.

3.1. Spatio-temporal Correlation in Video Sequence

Because of the nature of video, a frame sequence is chronologically tied to describe a story. Each image frame is sampled to capture a moment, forming a coherent relation with the adjacent frames. Spatio-temporal correlation indicates that these frames have a similar appearance and occur chronologically. We apply GMM to estimate the probability density of composition objects in individual frames using their feature distribution (e.g., colour, mfcc, and image intensity) [11]. The likelihood of spatial connection between two frames can be calculated using MLE (maximum likelihood estimation).

Technically, given a sequence of video segment \(X\) with frame length \(n\), \(d\)-dimensional feature distribution \(x_i\) for frame \(i\) is modelled by GMMs with \(A\) components. \(A\) represents a predefined number for frame composition such as the number of main objects found in the scene. Spatial information \(\Phi\) may be introduced to track the object location. For example, location of colour pixels can be specified in order to model the RGB colour distribution. The proximity of two frames is represented by the similarity of their appearance rather than global colour distribution.

The probability density for a video segment \(X\) is

\[
f(X) = \prod_{i=1}^{n} f(\Phi \oplus x_i | \Theta_{\Phi \oplus x_i})
\]

where \(\oplus\) implies concatenation meaning that new dimension is added to original data. \(\Theta_{\Phi \oplus x_i}\) is a set of GMM parameters derived from \(x_i\) with spatial information \(\Phi\). The EM (expectation maximisation) algorithm is employed to estimate the parameters [11]. The conditional distribution for \(x_i\) is defined as

\[
f(\Phi \oplus x_i | \Theta_{\Phi \oplus x_i}) = \sum_{a=1}^{A} \alpha_a f(\Phi \oplus x_i | \theta_a) = \sum_{a=1}^{A} \alpha_a f(x_i | \theta_a)
\]

where \(\alpha_a\) and \(\theta_a\) represent a priori probability and the \(a\)th component of \(\Theta_{\Phi \oplus x_i}\) with the mean \(\mu_a\) and the covariance matrix \(\Sigma_a\). 

\[
f(x_i | \theta_a) = \left(\frac{2\pi}{|\Sigma_a|}\right)^{\frac{d}{2}} \exp \left\{ -\frac{(x_i - \mu_a)^T \Sigma_a^{-1} (x_i - \mu_a)}{2} \right\}.
\]

Finally the spatio-temporal correlation \(\delta_{xx}\) of a sequence \(X\) is derived using MLE on the conditional distribution of \(X\):

\[
\delta_{xx} = \{\hat{\delta}_i, \hat{\delta}_j\}_{i=1 \ldots n} = \sum_{i=1}^{n} \sum_{j=1}^{n} f(x_i | \Theta_{x_j}). \tag{1}
\]

3.2. Manifold Learning for Spatio-temporal Synchronisation

Let \(X = \{x_1, \ldots, x_n\}\) and \(Y = \{y_1, \ldots, y_m\}\) represent two feature vectors derived from repetitive video sequences with length \(n\) and \(m\) respectively. Instances \(x_i\) and \(y_i\) are both \(d\)-dimensional vectors. The spatio-temporal correlation \(\delta_{xx}, \delta_{yy}\) are calculated as earlier. In order to estimate synchronisation for \(X\) and \(Y\), \(\delta_{xy}\) is additionally derived as follows:

\[
\delta_{xy} = \{\hat{\delta}_i, \hat{\delta}_j\}_{i=1 \ldots n, j=1 \ldots m} = \sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i | \Theta_{y_j}). \tag{2}
\]

We model a manifold embedding as a transformation \(T\) of high-dimensional data in terms of correlation \(\delta_{xx}, \delta_{yy}\) and \(\delta_{xy}\) into embedded configuration \(D\). The synchronisation is then generated as the optimal distance graph; each node represents a time instance on video sequences and an edge refers to spatio-temporal relationship between nodes:

\[
T : \{\delta_{xx}, \delta_{yy}, \delta_{xy}\} \rightarrow D. \tag{3}
\]

The function \(T\) is obtained by eigen decomposition of the integration of intra and inter-sequence correlation. We aim to minimise the loss function \(L_{projection}\):

\[
L_{projection} = ||\delta - T(\{\delta_x, \delta_y, \delta_{xy}\})||
= ||\delta - T(\delta)||
= ||\delta - (Q\wedge Q^T)||
= ||\delta - (Q^+\wedge \frac{1}{2})||. \tag{4}
\]
Fig. 2: Average precision and recall with $k$-nearest neighbours for MDS, LLE, Isomap and $k$-consecutive neighbours for Isomap-ST (i.e., spatio-temporal extension of Isomap). The operating points are $k = 7, 10, 15, 20, 25, 30, 35, 40$.

where $Q$ and $\Lambda$ are the eigenvectors and the eigenvalues of $\delta_\gamma$. To optimise the embedded representation, $\Lambda_+$ contains the $\epsilon$ largest eigenvalues in $\Lambda$ along the diagonal and $Q_+\Lambda_+$ is the square root of $\epsilon$ columns of $Q$ [12].

We apply Isomap to integrate the spatio-temporal correlations $\delta_{xy}$, $\delta_{yx}$, and $\delta_{xx}$ into a new correlation matrix $\delta_\gamma$. The method recalculates shortest distances along all frames in order to ensure that optimal related neighbors are closed by. The procedure is outlined below:

1. Given intra- and inter-correlation of two sequences, construct a neighbourhood graph consisting of nodes and edges. Each node represents a time instance of the frame sequence. An edge forms connection between two instances if they are possibly related. Each edge weights the similar probability of comparative instances.

For each sequence, $k$ consecutive frames are connected in order to form temporal correlation within the sequence. Inter-synchronisation is created by forming an edge to each other’s frames whose distance is inclusively lied on their pre-calculated temporal correlation.

2. Calculate shortest paths between nodes (i.e., frames) in the graph using Dijkstra’s shortest path algorithm [7]. The geodesic distance between all nodes in the graph is calculated, forming a new correlation matrix $\delta_\gamma$ (a matrix of pairwise geodesic distances).

4. EXPERIMENTS

The spatio-temporal extension of Isomap, or simply Isomap-ST, was experimented using MPEG-1 rushes video within the NIST TRECVID evaluation framework [2]. Six pairs of visually repetitive shots with the average length of two minutes were selected. The frame rate was 25 frame per second, however we processed 1 frame in every 25, practically in the rate of 1 frame per second. Each frame consisted of $288 \times 352$ pixels, from which RGB spatial colour features were extracted. The number of Gaussians in the mixture was set to ten.

The evaluation was made by human judges. We employed inclusion of video summary provided by NIST as our descriptive list of repetitive segments [13]. Three judges were asked to study the description and identify the location of segments related to the given detail. The position of all repetitions was utilised as a groundtruth. The accuracy for repetitive video synchronisation were measured based on their visual similarity to the groundtruth.

Two classical measures for information retrieval, precision and recall, were applied for evaluation of the video synchronisation performance. Figure 2 presents the comparison of Isomap-ST against other existing embedding methods; MDS, LLE, and the conventional Isomap. Individual performances were evaluated with varying value of $k$. The figure shows that the Isomap-ST captures the synchronisation of nearly repetitive sequences better than the rest of techniques.

For Isomap and LLE, $k$ was selected during graph construction. They would change the value when calculating shortest paths. Isomap-ST, on the other hand, adapted the value at first during the graph construction. Although MDS had no graph construction, all fixed $k$-nearest neighbours were selected. There was no great difference between MDS, LLE and Isomap for embedding two repetitive manifolds. Isomap-ST captured synchronisation well when $k$ was around 20.

Figure 3 illustrates the relation over two nearly repetitive sequences in rushs video, identified by MRS044499 in TRECVID. It is observed that Isomap-ST developed spatio-temporal relationship within and between sequences, resulting in clear boundaries in the figure. A sample frame sequence in MRS044499 is shown in Figure 4.

Although the sample was visually similar in segments (e) and (f), Isomap-ST could distinguish them using integrated context of spatial and temporal similarity (Figure 5a). The conventional Isomap (Figure 5c) failed to capture the synchronisation due to their spatial sim-
Fig. 4: Selection of consecutive frames in rushes video, chronologically ordered by alphabet (a) to (g). The segments were extracted from video data identified by MRS044499 in TRECVID.

Fig. 5: Synchronised map in 2- and 3-dimension at $k = 20$: Isomap-ST (top) vs conventional Isomap (bottom). Graphs on the left and the right illustrate the intra- and inter-synchronisation for two nearly repetitive sequences in MRS044499. The Magenta and blue maps represent two different sequences, while the inter-synchronisation is presented by yellow lines.

5. CONCLUSIONS AND FUTURE WORKS

The previous techniques for video synchronisation required frame-to-frame matching with camera calibration and object tracking. We have investigated a new approach to recovering frame relation between two nearly repetitive sequences using manifold embedding and a probabilistic model GMM. The experimental results indicated that the spatio-temporal extension of Isomap performed better than other conventional manifold embedding schemes. In the experiments, however, the value for $k$ was fixed before calculating the shortest paths, potentially including unnecessary neighbours. The performance may improve if $k$ can be adaptively calculated. We also plan to investigate the extension on multiple sequences aided by multimodal information such as audio, speech and text, which may further improve the accuracy. The contribution of this work may be applied to other applications on time series data such as pattern recognition, analysis, and visualisation.

6. REFERENCES