An EM Algorithm for Video Summarization, Generative Model Approach

Xavier Orriols, Xavier Binefa

Computer Vision Center, Universitat Autònoma de Barcelona
08193 Bellaterra, Spain, {xevi,xavierb}@cvc.uab.es

Abstract

In this paper, we address the visual video summarization problem in a Bayesian framework in order to detect and describe the underlying temporal transformation symmetries in a video sequence. Given a set of time correlated frames, we attempt to extract a reduced number of image-like data structures which are semantically meaningful and that have the ability of representing the sequence evolution. To this end, we present a generative model which involves jointly the representation and the evolution of appearance. Applying Linear Dynamical System theory to this problem, we discuss how the temporal information is encoded yielding a manner of grouping the iconic representations of the video sequence in terms of invariance. The formulation of this problem is driven in terms of a probabilistic approach, which affords a measure of perceptual similarity taking both learned appearance and time evolution models into account.

1. Introduction

Browsing and retrieval by content in video data-bases is becoming a relevant field in Computer Vision and Multimedia Computing. This fact goes in accordance with the increasing developments in digital storage and transmission. In addition to this, the wide range of applications in this framework, such as advertising, publishing, news and video clips, points out the necessity for more efficient organizing techniques [2, 7].

In this paper, we focus on two important subjects in this area, video preview and summarization, and, which make feasible a quick intuition of the evolution, (under a low streaming cost), of higher-level perceptual structures, such as stories, scenes or pieces of news. That fact becomes relevant for low bandwidth communication systems. Expressing a video sequence in terms of a few representative images permits a continuous media to be seekable. Besides, the summarizing ability of a story will depend on the specific choice of key-frame set. Currently, the standard approach for keyframes selection, as indicators of the content of video, is to choose certain images that belong to the video sequence, which usually correspond to the beginning and the end of clips. However, considering that editors, authors and artists utilize camera operations to communicate some specific intentions, this standard key-frame selection may presents the risk of losing semantic information.

For this reason, our purpose is to present a compact and perceptually meaningful representation that preserves the subjective approach, i.e. the semantics, given by actions and camera operations in the evolution of a video sequence. The model to extract this new set of iconic representative image -like data structures is based on an application of Linear Dynamical System and Lie’s group theories, which are our support to define temporal symmetries and invariances. In this framework, the temporal information is encoded in an infinitesimal generator matrix, which defines different types of behaviors in the evolution of an image sequence. We use this distinct sort of contributions to give, in addition, a grouping inside the summarized representation.

The formulation of this problem is driven in terms of a probabilistic approach. Appearance representation and time evolution between consecutive frames are introduced in a generative model framework. First, a feature space is built through Probabilistic Principal Component Analysis (PPCA) [1], since this technique allows us to codify images as points capturing the intrinsic degrees of freedom of the appearance, and at the same time, it yields compact description preserving semantics and perceptual similarities [9, 6, 5]. Subsequently, we present a generative dynamical model for the estimation of the curve’s behavior that the sequence of images describe in this subspace of principal features. Authors in [8] introduced previously this dynamical model in a neural network framework. However, what distinguishes our work is that we embed it into a latent variable model, providing an EM algorithm for its estimation. This fact avoids undesirable problems such as when it comes to manually assigning the update steps of gradient descents techniques. Furthermore, the presented latent variable model allows a conjugation of both semantic and temporal representations. This affords a measure of perceptual similarity taking both learned appearance and time evolution sub models into account. Indeed, this probabilistic framework allows determining whether two consecutive images are in accordance with the learned dynamical model. This fact has an important significance when it comes to assign some boundaries to a sequence of frames.
The outline of this paper is as follows: first, we introduce a review on Linear Dynamical Systems. The aim of this is to present the key points on the interpretation of the temporal appearance codification and how this information can be extracted. Subsequently, in section 3, an appearance probabilistic framework for time symmetry estimation is introduced in terms of latent variable models. Section 4 shows the experimental results in order to see this framework applied to real image problems. Section 5 presents the summary and conclusions. Finally, the appendix gives a detailed explanation of the developed EM algorithm for the dynamical model estimation.

2. On underlying symmetries

Consider a sequence of frames $\mathcal{F} = \{\phi_0, \ldots, \phi_N\}$ that are represented as vectors. Each vector corresponds to an image read in lexicographic order belonging to a subset of real numbers $\mathcal{S} \subset \mathbb{R}^d$. Since images are obtained from a temporal sequence, order takes significant relevance where transition between two consecutive frames is achieved as a transformation from the previous one to the next one. Suppose that this transformation can be parameterised by a single real number $\theta \in \mathbb{R}$, which gives us a notion of time over the whole sequence. Therefore, when $\theta$ tends to zero, the associated transformation is the identity, recovering the initial image $\phi_0$. In a first approximation order, the relation between an image $\phi_0$ and a near one transformed $\phi(\theta \theta)$ can be expressed as: $\phi(\theta \theta) \approx (1 + \theta \theta G) \phi_0$. So, a macroscopic transformation $T(\theta)$ can be built in terms of concatenating infinitesimal transformations, dividing the parameter $\theta$ in $M$ parts and making $M \to \infty$:

$$\phi(\theta) = \lim_{M \to \infty} \left(1 + \left(\frac{\theta}{M}\right) G\right)^M \phi_0 = e^{\theta G} \phi_0 \quad (1)$$

Equation (1) is related to the study of a trajectory near a fixed point in $\mathcal{S}$ described by a linear dynamical system:

$$\dot{\phi} = G \phi \quad (2)$$

The basic idea considering a trajectory in $\mathbb{R}^d$, formed by a sequence of images $\mathcal{F} = \{\phi_0, \ldots, \phi_N\}$, is to understand what its underlying appearance invariance. From a geometric point of view invariance is defined as follows:

**Definition 2.0.1** Let $\mathcal{S} \subset \mathbb{R}^d$ be a set, then $\mathcal{S}$ is said to be invariant under the vector field $\dot{\phi} = T(\phi)$ if for any $\phi_0 \in \mathcal{S}$ we have $\phi(\theta, \phi_0) \in \mathcal{S}$ for all $\theta \in \mathbb{R}$.

Furthermore, we can see that the information available in the temporal evolution of a sequence of frames is encoded in the matrix $G$ under this linear model. The goal is to find how this information can be extracted. To this end, in the following section we describe the geometrical meaning of that matrix $G$, as well as, the behavior that follow the solutions of eq. (2) from the analysis of the internal structure of the linear system.

2.1. Geometrical Point of View of Dynamical Systems

In order to give an intuitive idea of the behavior of the solutions in eq. (2), we focus on an analysis of the orbit structure near fixed points. In eq. (1), a macroscopic transformation was built by considering a continuous process with incremental changes in the evolution parameter $\theta$. This type of transformations form a one-parameter Lie group, which satisfies the following differential equation:

$$\frac{dG(\theta)}{d\theta} = G T(\theta)$$

that corresponds to a generalization of the plane rotation and translation groups. The matrix $G$ is called infinitesimal generator or action of the group. Lie’s group theory applied to Computer Vision is not new. In order to get an insight into this framework, we recommend [4], where a comprehensive view of its applications is developed.

Besides, the evolution described by (2) is a particular case of dynamical systems. Indeed, it corresponds to consider a linearization a system of differential equations:

$$\dot{\phi} = F(\phi) \rightarrow \dot{\phi} = DF(\phi) |_{\phi(0)} \phi \rightarrow \dot{\phi} = G \phi$$

where $G = DF(\phi) |_{\phi(0)}$ with $\phi \in \mathbb{R}^d$. This development is carried out through the analysis in the vicinity of a certain fixed point $\phi(0)$ at $\theta = 0$. In the following sections we show how this approximation can be assumed embedding the estimation problem into a probabilistic framework. This fact affords a measure of likelihood that determines whether a set of consecutive points $\{\phi(\theta_n)\}$ are as a result of a certain transformation of this type.

The Linear Dynamical Systems theory shows how to extract information of the system by means of an eigenvector analysis of the infinitesimal generator $G$. The starting point is that the $\mathbb{R}^d$ space can be represented as a direct sum of three subspaces defined in terms of a set of (generalized) eigenvectors: $E^s = \text{span} \{e_1, \ldots, e_s\}$, $E^u = \text{span} \{e_{s+1}, \ldots, e_{s+u}\}$ and $E^c = \text{span} \{e_{s+u+1}, \ldots, e_{s+u+c}\}$. The first set of eigenvectors $\{e_1, \ldots, e_s\}$ corresponds to the eigenvalues of $G$ having negative real part, the second set are the eigenvectors $\{e_{s+1}, \ldots, e_{s+u}\}$ whose corresponding eigenvalues have positive real part, and $\{e_{s+u+1}, \ldots, e_{s+u+c}\}$ correspond to the eigenvalues of $G$ with zero real part. These subspaces are called, stable subspace $E^s$, unstable subspace $E^u$ and center subspace $E^c$ respectively, and $s + c + u = d$.

These spaces are an example of invariant subspaces, since solutions of eq. (2) with initial conditions entirely contained in either $E^s$, $E^u$ or $E^c$ must remain in that particular subspace for all values of $\theta$ (time) according to the definition 2.0.1.

In order to see the meaning of the eigenvalues of $G$, let us consider the following example. A curve in $\mathbb{R}^4$ is built...
3 subspaces, \( E^c = \text{span} \{e_1, e_2\}; \{i\omega, -i\omega\} \), \( E^u = \text{span} \{e_3\}; \lambda_1 \) and \( E^s = \text{span} \{e_4\}; \lambda_2 \). Taking \( \phi(0) = [x_1(0), x_2(0), x_3(0), x_4(0)] \) as initial condition, the parametric equation of the orbit, i.e. the solution of eq. (2) in this particular case is:

\[
\phi(\theta) = T e^{\theta \mathbf{\Omega}} T^{-1} \phi(0) = \begin{vmatrix}
\cos \omega \theta & \sin \omega \theta & 0 & 0 \\
-\sin \omega \theta & \cos \omega \theta & 0 & 0 \\
0 & 0 & e^{\lambda_1 \theta} & 0 \\
0 & 0 & 0 & e^{\lambda_2 \theta}
\end{vmatrix}
T^{-1} \phi(0)
\]

(3)

From this solution, we deduce that \( E^c = \text{span} \{e_1, e_2\} \) (purple and black axis directions in fig. 1 (f)) is an invariant subspace that generates a closed orbit, \( E^u = \text{span} \{e_3\} \) (red axis direction in fig. 1(b)) is an invariant subspace of solutions that decay to zero as \( \theta \to -\infty \), and \( E^s = \text{span} \{e_4\} \) (green axis direction in fig. 1(d)) is the third invariant subspace of solutions that decay to zero as \( \theta \to \infty \). For instance, consider an initial condition like the following:

\[
T^{-1} \phi(0) = \begin{pmatrix}
x_1(0) \\
x_2(0) \\
0 \\
0
\end{pmatrix}
\]

Therefore, orbit obtained by means of (3) remains in \( E^c = \text{span} \{e_1, e_2\} \) for all possible values of the time parameter \( \theta \), and that fact is in accordance with the definition 2.0.1.

With this reference to the analysis of the solutions of linear dynamical systems, we see that the information, which is encoded in the infinitesimal generator \( G \), is straightforward understandable through its eigenvalues and eigenvectors. This internal structure analysis not only allows the selection of a new representation for the images evolution, which is based in the modes of \( G \) (eigenvectors), but also yields a manner of grouping the different principal directions of \( G \) distinguishing the subspaces that they span in terms of stability, i.e., \( E^u, E^s \) and \( E^c \).

3. Appearance Based Framework for Time Symmetry Estimation

The previous example was performed in order to illustrate that a sequence of points that follow a certain temporal evolution can be described in terms of some privileged directions, which are indicative of the behavior of the curve. The aim of this is to apply linear dynamical system theory to temporal correlated sequences of images. To this end, we need to define a feature space where the images can be represented as points. Subsequently, the goal is to have an estimation of the evolution process of these images. The aim of this is to extract the temporal information encoded in the infinitesimal generator in order to present a reduced number of lower dimensional subspaces. Now, we can see that the \( \mathbb{R}^4 \) space is decomposed in terms of stability, i.e., \( E^u, E^s \) and \( E^c \).
of images that are capable of summarizing semantically the whole sequence.

In this section, we present a generative model which defines appearance representation and time evolution between consecutive frames. This model involves jointly representation and evolution of appearance. In this case, temporal symmetry estimation is based on the fact that images belonging to a coherent sequence are also related by means of appearance representation.

First, the probabilistic formulation for appearance description is developed in terms of linear generative models through Probabilistic Principal Component Analysis (PPCA) [1]. Subsequently, the temporal appearance evolution is developed inside the linear generative model with the purpose of presenting a unified framework, where likelihood measure takes into account both appearance representation and evolution.

3.1. Appearance Representation Model

First of all, we need to define a space of features where images are represented as points. This problem involves finding a representation as a support for analyzing the temporal evolution. To address the problem of appearance representation, authors in [9, 6, 5] proposed Principal Component Analysis as redundancy reduction technique in order to preserve the semantics, i.e. perceptual similarities, during the codification process of the principal features. The idea is to find a small number of causes that in combination are able to reconstruct the appearance representation. These small numbers of causes are taken as the basis for the feature space. Besides, Tipping et. al. [1] embedded PCA into a Linear Generative Model framework in order to capture the intrinsic degrees of freedom of the object category model as well as to give an inherent likelihood measure to the learned object category. Generative models are a causal approach to describe the underlying phenomena that generates the complexity of observed data (images).

One of the most common approaches for explaining a data set is to assume that causes in linear combination:

\[ t = Wx + \mu + e \]

where \( x \in \mathbb{R}^q \) (our chosen reduced representation, \( q < d \)) are the causes (latent variables), \( W \) is an orthogonal matrix which rotates the data \( t \), \( \mu \) corresponds to the sample mean and \( e \) is some noise. This causal approach leads to define a joint distribution \( p(t, x) \) over visible \( \{t\} \) and hidden variables \( \{x\} \), the corresponding distribution \( p(t) \) (similarity measure) for the observed data is obtained by marginalization: \( p(t) = \int p(t | x)p(x)dx \), where \( p(t | x) \) defines the causal connection between the observations \( \{t\} \) and the latent variables \( \{x\} \), and it is associated to the noise distribution as follows:

\[
p(t | x) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp \left\{ -\frac{1}{2\sigma^2} \| t - \mu - Wx \|^2 \right\}
\]

and the corresponding similarity measure given the model:

\[
p(t) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp \left\{ -\frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu) \right\}
\]

where \( \Sigma = WW^T + \sigma^2 I \). The prior knowledge on latent variables is expressed in \( p(x) \). This density function takes the form of a Gaussian distribution with zero mean and identity covariance matrix: \( \mathcal{N}(0, I) \). Therefore, it is said that the causes are mutually independent in terms of a second order statistics. The main goal is to find the parameters that maximize the joint observed data distribution i.e. the best description under the specific generative model. After considering the temporal model, the algorithm to estimate latent variables and parameters is introduced in a unified framework for appearance representation and evolution.

3.2. Appearance Temporal Evolution Model

Given a suitable basis to describe appearance, temporal symmetry can be analyzed in terms of this representation. An image \( t \) corresponds to a point \( x \) in the latent space, \( S \). On the other hand, equation (1) can be interpreted in terms of a generative model, where an image description in latent space \( x_{n+1} \) is obtained by the action of the infinitesimal generator (matrix \( G \)), a certain quantity \( \theta_n \), on a previous one \( x_n \).

Symmetry learning is based on observations, more specifically, in a sequence of ordered images. So, is feasible to consider that observations are obtained with a certain additive noise. The generative equation takes the following form:

\[
x(\theta) = e^{\theta G} x(0) + r \quad (4)
\]

where \( r \) is a \( \theta \)-independent noise process. According to the infinitesimal approximation \( e^{\theta G} \sim 1 + \theta G \), eq. (4) yields:

\[
x(\theta) = (1 + \theta G) x(0) + r \\
\Delta x(\theta) = \theta G x(0) + r \quad (5)
\]

where \( \Delta x(\theta) \equiv x(\theta) - x(0) \). For the isotropic noise model case \( r \sim \mathcal{N}(0, \beta^2 I) \), the probability distribution over the transformations \( \Delta x \)-space for a given image \( x \in S \) and step parameter \( \theta \) corresponds to:

\[
P(\Delta x \mid x, \theta, G, \beta^2) = \frac{1}{(2\pi\beta^2)^{d/2}} \exp \left\{ -\frac{1}{2\beta^2} \| \Delta x - \theta G x \|^2 \right\}
\]

The prior distribution over the latent variables \( \theta \) is assumed to be Gaussian with unit variance, so \( \theta \sim \mathcal{N}(0, I) \). Therefore, the corresponding similarity measure for temporal transformations in latent space \( S \) is obtained by marginalization:

\[
P(\Delta x \mid x, G, \beta^2) = \int P(\Delta x \mid x, \theta, G, \beta^2)P(\theta) d\theta = \frac{1}{(2\pi)^d \sqrt{\det C}} \exp \left\{ -\frac{1}{2} \Delta x^T C^{-1} \Delta x \right\}
\]
where \( C = G x x^T G^T + \beta^2 I \). The similarity measure eq. (6) evaluates the likelihood of a transformation \( \Delta x \) between two points, for a given \( x_n \) to a following one \( x_{n+1} \), w.r.t. a learned model \( \{ G, \beta^2 \} \). These points \( \{ x_n, x_{n+1} \} \) are a representation of two images \( \{ t_n, t_{n+1} \} \) for a certain instance of the appearance model \( \{ W, \mu, \sigma^2 \} \). Indeed, this probabilistic framework allows determining whether two consecutive images are in accordance with the learned dynamical model. This fact has an important significance when it comes to assign some boundaries to a sequence of frames.

3.3. Maximum Likelihood Estimation

At this point, the problem is centered on parameter estimation, which, in practice, will be given by data observations. This leads to consider the problem of incomplete data. For this purpose, Dempster et al. (1977) [3] use the EM algorithm, where each observation \( t_n \) (image) is associated to an unobserved state \( s_n = \{ x_n, \theta_n \} \), and the main goal is to determine which component generates the observation. In this sense, the unobserved states can be seen as missing data and therefore the union of observations \( t_n \) and \( s_n \) is said to be complete data, \( y_n = \{ t_n, s_n \} \). In this way, for a given set of observations \( \{ t_1, \ldots, t_N \} \) the likelihood measure to be maximized is the Complete-log-Likelihood, i.e.:

\[
\mathcal{L}(y_1, \ldots, y_N \mid \Omega) = \log \{ p(t_1, \ldots, t_N; s_1, \ldots, s_N \mid \Omega) \} \tag{7}
\]

where \( \Omega \) represents the model parameters. Although both parameters and latent variables are unobserved, the difference is that latent variables are presumed to be evaluated once every observation, that is there is a latent \( s_n \) for each observation \( t_n \). Furthermore, the noise model offers smoothness, therefore, this approach differs from regression-based methods, in the way that the goal is to estimate the data density, and leads to reduce the overfitting. The following table shows the parameters \( \Omega \) that are involved in the model, and the latent variables related to \( s_n \):

<table>
<thead>
<tr>
<th>Model</th>
<th>Generative Mapping</th>
<th>Parameters, ( \Omega )</th>
<th>( s_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>App. Rep.</td>
<td>( t_n = W x_n + \mu + \epsilon_n )</td>
<td>( W, \sigma^2, \mu )</td>
<td>( x_n )</td>
</tr>
<tr>
<td>Time Sym.</td>
<td>( \Delta x_n = \theta_n G x_n + r_n )</td>
<td>( G, \beta^2 )</td>
<td>( \theta_n )</td>
</tr>
</tbody>
</table>

Following the assumption that appearance representation depends only on data observations, the ML estimation for the appearance parameters is given in a closed form solution as it is developed in [1]:

\[
W = U_q (\Lambda_q - \sigma^2 I)^{1/2} R; \quad \sigma^2 = \frac{1}{d - q} \sum_{j=q+1}^{d} \lambda_j
\]

where \( U_q \) are the first \( q \) eigenvectors of the data set covariance matrix, \( \Lambda_q \) is a diagonal matrix with the corresponding first \( q \) eigenvalues \( \lambda_i \) \( 1 \leq i \leq N \) and \( R \) is an arbitrary rotation matrix.

In order to estimate the appearance dynamics we utilize an EM algorithm, which is detailed in the Appendix A. This is basically a two steps procedure: Expectation and Maximization of a likelihood function. The Expectation step requires a third operation, which in the latent variable model can be added to the pair of learning and model selection: inference. This refers to estimation of value of latent variables \( s_n \) given known parameters \( \Omega \) and observations \( t_n \).

The introduced model shows a hierarchical structure between observed images \( t_n \) and latent variables \( x_n, \theta_n \). First, images are obtained to build the appearance representation, and secondly, taking advantage of a reduced appearance basis, data evolution is estimated. Inference in this framework is a simple matter of the application of Bayes’ rule:

1. Inferring latent variables related to appearance:

\[
p(x \mid t, \Omega) = \frac{p(x)p(t \mid x, \Omega)}{p(t \mid \Omega)} \tag{8}
\]

2. Inferring latent variables related to temporal evolution, given appearance latent variables inferred and their corresponding transformations computed:

\[
p(\theta \mid x, \Delta x, \Omega) = \frac{p(\theta)p(\Delta x \mid \theta, x, \Omega)}{p(\Delta x \mid x, \Omega)} \tag{9}
\]

Therefore, for each image \( t_n \) the computation of its corresponding coordinates in latent space \( x_n \) is given by means of the Maximum a Posteriori (MAP) in eq.(8):

\[
x = \arg\max_{x'} p(x' \mid t, \Omega) \tag{10}
\]

Once, images are expressed in latent space coordinates, the computation of the best estimated transformation parameter \( \theta_n \) is also done through the MAP in eq.(9):

\[
\theta = \arg\max_{\theta'} p(\theta' \mid x, \Delta x, \Omega) \tag{11}
\]

Under gaussian assumptions for noise models and prior knowledges, the posterior means \( < x \mid t > \) and \( < \theta \mid x, \Delta x > \) correspond to the MAP for each distribution. In the appendix we show the explicit forms for these posterior probabilities, and an EM algorithm for the temporal parameters estimation is introduced.

4. Experimental Results

In this section we apply the introduced appearance evolution model to the extraction of semantically meaningful image-like data structures. We analyze two sequences: one focuses on the evolution of an action (fig. 2(a)), and the other corresponds to a camera operation, fig. 3(a). The estimation process is common for both cases, but first it is necessary to build the appearance representation by means of the ML for PPCA. Afterwards, utilizing the posterior probability eq. (8), the new coordinates for the images are
Figure 2: (a) Original sequence of frames. Reconstructed sequence (b) with 70.45% of reconstruction quality using 2 principal components of appearance (e) and (f) and a $2 \times 2$ $G$ matrix whose eigenvectors correspond to the images (c) and (d).

Figure 3: (a) Original sequence of frames. Reconstructed sequence (b) with 88.17% of reconstruction quality using 4 principal components of appearance (e) and a $4 \times 4$ $G$ matrix whose eigenvectors correspond to the images (c) and (d).
computed though MAP. We use these new appearance representations to estimate the dynamical model by means of the introduced EM algorithm (see appendix). Once $G$ is estimated we compute the iconic representation, by means of an eigenvector analysis of the infinitesimal generator. These eigenvectors are back-projected to the image space in order to have an expression of them as images.

The first sequence, fig. 2 (a), is represented by a 2-dimensional appearance eigenspace with a 70.45% of reconstruction quality. The appearance eigenvectors represented as images are shown in fig. 2(e) and (f). These express the variations between the mean or prototypical appearance of the object. However, such a prototype is not an appropriate representative sample, since the concept of “face” is not as perceptible as in figs. 2(c), (d). Performing the estimation of the dynamical model $\{G, \beta^2\}$, we have the corresponding generator of the transformation as a $2 \times 2$ matrix whose modes correspond to the images fig. 2 (c), (d). They are presented like real images, however, they are not directly obtained from the sequence, i.e., like selecting the first frame and last one. The significant issue here is that this temporal information can be used to reconstruct the video sequence (fig. 2 (b)) by means of eq. (5) and using just only the $2 \times 2$ matrix, the appearance basis and the time step $\theta_n$ of the evolution, (see fig. 4(a)).

The purpose of this second sequence (fig. 3) is to show that the appearance evolution model does allow us to keep perceptually the camera operation in the new iconic representation. Given that editors, authors and artists communicate some specific intentions with certain camera operations, we notice that this information remains in the sum-

mation. We address the problem of characterizing key-frames basing partitions on a visual appearance information criterion and, at the same time, conjugating semantic and temporal representations. This fact, not only allows embedding in a more numerical tractable framework the video retrieval, but also yields a new approach to extract underlying information from temporal evolution of sequences. A suitable selection of these basic perceptual units allows the transformation of a continuous temporal data structure into a discrete meaningful one, where the intention is that the semantics remains preserved. The choice of an appropriate representation for the data takes a significant relevance when it comes to deal with symmetries, since these usually imply that the number of intrinsic degrees of freedom in the data distribution is lower than the coordinates used to

Figure 4: (a) Time step $\theta_n$ inferred values for each image of the sequence fig. 2(a). (b) Inferred $\theta_n$ values for the sequence fig. 3(a)

5. Summary and Conclusions

As an alternative to standard key-frames selection, in this paper we propose a Bayesian framework for video summarization. We address the problem of characterizing key-frames basing partitions on a visual appearance information criterion and, at the same time, combining semantic and temporal representations. This fact, not only allows embedding in a more numerical tractable framework the video retrieval, but also yields a new approach to extract underlying information from temporal evolution of sequences. A suitable selection of these basic perceptual units allows the transformation of a continuous temporal data structure into a discrete meaningful one, where the intention is that the semantics remains preserved. The choice of an appropriate representation for the data takes a significant relevance when it comes to deal with symmetries, since these usually imply that the number of intrinsic degrees of freedom in the data distribution is lower than the coordinates used to
represent it. Indeed, this means that the problem has been reduced to a lower dimensional one. Therefore, using both topics, the decomposition into basic units and the change of representation, transform a complex problem into a manageable one. These simplifications of the estimation problem rely on a proper mechanism of combination of those primitives (appearance eigenvectors) in order to give an optimal description of the global complex model.

A EM algorithm for Maximum Likelihood Estimation

In the Expectation-Maximization (EM) algorithm for symmetry estimation, we consider the latent variables \( \theta_n \) to be "missing". If their value were known, the estimation of \( G \) would be straightforward from equation (5) by applying standard least-squares techniques. This leads us to consider a joint distribution over visible and hidden variables, then the corresponding distribution for the observed data is obtained by marginalisation. Thus the goal is to find the parameters that maximize the joint observed data distribution i.e. the best description under a specific generative model. Assuming that we do not know for a given transformation \( \Delta x_n \), which value of \( \theta_n \) generated it, the joint distribution can be calculated through the expectation of the complete data log-likelihood.

From a Bayesian point of view, Maximum Likelihood estimation requires, in general, a two step procedure:

1. **Expectation** of latent variables \( \theta_n \) for a given observed \( \Delta x \). Posterior expectation given observed data for the complete log-likelihood is a way of computing an approximated predictive density.

2. **Maximisation** of complete log-likelihood from the model parameters \( G, \sigma^2 \). This is equivalent to minimizing the expectation of the loss function under that (approximated) predictive density.

**Expectation**

The expected complete-data log likelihood is given by:

\[
\mathcal{L} = \sum_{n=0}^{N-1} \left\{ \frac{1}{2 \sigma^2} | \Delta x_n - \theta_n G x_n |^2 + \frac{q}{2} \log \sigma^2 + \frac{< \theta_n^2 >}{2} \right\}
\]

which expanded yields:

\[
\mathcal{L} = \sum_{n=0}^{N-1} \left\{ \frac{1}{2 \sigma^2} | \Delta x_n |^2 + \frac{1}{2 \sigma^2} < \theta_n^2 > x_n^T G^T G x_n - \frac{1}{2 \sigma^2} < \theta_n > x_n^T G^T \Delta x_n - \right\}
\]

\[
- \frac{1}{2 \sigma^2} < \theta_n, > \Delta x_n^T G x_n + \frac{< \theta_n^2 >}{2} + \frac{N q}{2} \log \sigma^2
\]

where the sufficient statistics of the posterior distributions correspond to:

\[
< \theta_n > = M^{-1} x_n^T G^T \Delta x_n \quad (12)
\]

\[
< \theta_n^2 > = \sigma^2 M^{-1} + < \theta_n > < \theta_n > \quad (13)
\]

with \( M = \sigma^2 + x_n^T G^T G x_n \).

**Maximisation**

Differentiating equation (12) and setting the derivatives to zero, \( G \) and \( \sigma^2 \) are updated as:

\[
\hat{G} = \left[ \sum_n < \theta_n > \Delta x_n x_n^T \right] \left[ \sum_n < \theta_n^2 > x_n x_n^T \right]^{-1}
\]

\[
\hat{\sigma}^2 = \frac{1}{N \hat{q}} \sum_{n=1}^{N} \left\{ | \Delta x_n |^2 - 2 < \theta_n > x_n^T \hat{G}^T \Delta x_n + < \theta_n^2 > (\hat{G} x_n) \right\}
\]

These equations are iterated until the algorithm converges with a certain degree of tolerance.

References


