ISSN 2320-9984 04 >
9 772320 998015
Volume 1, Issue 4, December 2013 ISSN: 2320-9984 (Online)
International Journal of Modern Engineering & Management Research
Website: www.ijmemr.org

Application of Hierarchical EM Algorithm for Modeling Video in Dynamic Texture

Yogesh Singh Baghel
M. Tech. Research Scholar
Department of Computer Science & Engineering
Patel College of Science and Technology Bhopal
Bhopal (M.P.) [INDIA]
Email :yogesh.baghel2008@gmail.com

Prof. Hitesh Gupta
Assistant Professor
Department of Computer Science & Engineering
Patel College of Science and Technology Bhopal
Bhopal (M.P.) [INDIA]
Email : hitesh034@gmail.com

Abstract*—To observe the video sequencing from a linear dynamical system a dynamic texture of spatio-temporal generative model is used. This work studies the mixture of dynamic textures, a statistical model for an ensemble of video sequences that is sampled from a finite collection of visual processes, each of which is a dynamic texture. The proposed technique automatically determines the optimum number of clusters for modeling of video. The optimal number of clusters is obtained by using Hierarchical EM Algorithm. The proposed algorithm is evaluated on well known natural images and its performance is compared to other clustering techniques. Experimental results show the performance of the proposed algorithm producing comparable in modeling of video in dynamic texture.*

Keywords:*— Dynamic textures, expectation maximization, Kalman filter, bag of systems, video annotation, sensitivity analysis.*

1. INTRODUCTION

Colour images are more complex than gray scale images as instead of a single intensity value for a pixel, each pixel is usually denoted by three component values such as Red, Green and Blue. Clustering based methods are ideal to use for gray scale images can be easily extended to cope with higher dimensity, although, the increased dimensionality also leads to more computationally expensive process. Various segmentation techniques have been developed for image segmentation include Unsupervised Colour Textured Image Segmentation Using Cluster Ensembles and MRF Model [1,2], Determination of Number of Clusters in k-means Clustering and application in Colour Image Segmentation [3], Unsupervised Colour Image Segmentation based on Gaussian Mixture Model [4] etc likewise modeling motion as a spatiotemporal texture has shown promise in a wide variety of computer vision problems which have otherwise proven challenging for traditional motion representations such as optical flow. In particular, the dynamic texture (DT) model proposed in has demonstrated a surprising ability to abstract a wide variety of complex global patterns of motion and appearance into a simple spatiotemporal model. The dynamic texture is a probabilistic generative model, defined over space and time that represents a video (i.e., spatiotemporal volume) as the output of a linear dynamical system (LDS).

Natural colour images are particularly noisy due to the environment they were produced. Therefore, it is hard to develop a robust and faithful unsupervised technique for automatic determination of number of objects in a colour image. Although there are a few existing approaches for unsupervised colour image segmentation, none of them has been

found robust in all situation. Initially we tried SNOB[5], a Minimum Message Length (MML) based unsupervised data clustering approach to address this problem. In this paper, we address the problem of clustering dynamic texture models, i.e., clustering linear dynamical systems. Given a set of DTs (e.g., each learned from a small video cube extracted from a large set of videos), the goal is to group similar DTs into *K* clusters, while also learning a representative DT "center" that can sufficiently summarize each group. This is analogous to standard *K*-means clustering, except that the datapoints are dynamic textures instead of real vectors. A robust DT clustering algorithm has several potential applications in video analysis, including [a] Hierarchical clustering of motion, [b] Obtaining fast video retrievals [c] Generation of DT codebook. [d] Weakly supervised learning by semantic video annotation. Application of hierarchical estimation DT clustering can also serve as an effective method for learning DTs from a large dataset of video.

The parameters of the LDS lie on a non-euclidean space (nonlinear manifold), the K-means algorithm cannot be clustered directly. First embeds the DTs into a euclidean space using nonlinear dimensionality reduction (NLDR), and then performs K-means on the low-dimensional space to obtain the clustering. While this performs the task of grouping the DTs into similar clusters, [6,7] is not able to generate novel DTs as cluster centers. These limitations could be addressed by clustering the DTs' parameters directly on the nonlinear manifold, e.g., using intrinsic mean-shift [8,9] or LLE [10]. However, these methods require analytic expressions for the log and exponential map on the manifold, which are difficult to compute for the DT parameters. In the present study the demonstration of the efficacy of the HEM clustering algorithm has been shown for DTs on several computer visions problems. Initially the hierarchical clustering of video textures has been performed to show that HEM groups perceptually similar motion together. In the next step HEM method has been used to learn DT mixture models for semantic motion annotation, based on the supervised multiclass labeling (SML) framework.

DT annotation models are learned efficiently from weakly labeled videos by aggregating over large amounts of data using the HEM algorithm. Lastly codebooks have been generated with novel DT codeword for the bag-of-systems motion representation, and demonstrate improved performance on the task of dynamic texture recognition [11]. The HEM algorithm for GMMs proposed in [12] has been employed in [13] to build GMM hierarchies for efficient image indexing and in [14] to estimate GMMs from large image datasets for semantic annotation. In this paper, we extend the HEM algorithm to dynamic texture mixtures (DTMs), where each mixture component is an LDS. In contrast to GMMs, the E-step inference of HEM for DTMs requires a substantial derivation to obtain an efficient algorithm due to the hidden state variables of the LDS.

2. OTHER STUDIES

The dynamic texture (DT) is a probabilistic generative model, defined over space and time, that represents a video as the output of a linear dynamical system (LDS). The DT model has been applied to a wide variety of computer vision problems, such as motion segmentation, motion classification, and video registration. In this paper, we derive a new algorithm for clustering DT models that is based on the hierarchical EM algorithm. The proposed clustering algorithm is capable of both clustering DTs and learning novel DT cluster centers that are representative of the cluster members, in a manner that is consistent with the underlying generative probabilistic model of the DT.

We then demonstrate the efficacy of the clustering algorithm on several applications in motion analysis, including hierarchical motion clustering, semantic motion annotation, and bag-of-systems codebook generation. [15] A dynamic texture is a spatio-temporal

generative model for video, which represents video sequences as observations from a linear dynamical system. This work studies the mixture of dynamic textures, a statistical model for an ensemble of video sequences that is sampled from a finite collection of visual processes, each of which is a dynamic texture. An expectation-maximization (EM) algorithm is derived for learning the parameters of the model, and the model is related to previous works in linear systems, machine learning, time- series clustering, control theory, and computer vision.

Through experimentation, it is shown that the mixture of dynamic textures is a suitable representation for both the appearance and dynamics of a variety of visual processes that have traditionally been challenging for computer vision (for example, fire, steam, water, vehicle and pedestrian traffic, and so forth). When compared with state-of-the-art methods in motion segmentation, including both temporal texture methods and traditional representations (for example, optical flow or other localized motion representations), the mixture of dynamic textures achieves superior performance in the problems of clustering and segmenting video of such processes[16]. A dynamic texture is a linear dynamical system used to model a single video as a sample from a spatio-temporal stochastic process.

In this work, we introduce the mixture of dynamic textures, which models a collection of videos consisting of different visual processes as samples from a set of dynamic textures. We derive the EM algorithm for learning a mixture of dynamic textures, and relate the learning algorithm and the dynamic texture mixture model to previous works. Finally, we demonstrate the applicability of the proposed model to problems that have traditionally been challenging for computer vision[17]. Dynamic textures are video sequences of complex nonrigid dynamical objects such as fire, flames, water on the surface of a lake, a flag fluttering in the wind, etc.

The development of algorithms for the analysis of such video sequences is important in several applications such as surveillance, where, for example, one wants to detect fires or pipe ruptures. However, the continuous change in the shape and appearance of a dynamic texture makes the application of traditional computer vision algorithms very challenging. Over the years, several approaches have been proposed for modeling and synthesizing video sequences of dynamic textures [18, 19], Among them, the generative model proposed by Doretto et al. [20], where a dynamic texture is modeled using a Linear Dynamical System (LDS), has been shown to be very versatile.

The LDS-based model has been successfully used for various vision tasks such as synthesis, editing, segmentation, registration, and categorization. In this paper, we are primarily interested in the problem of categorization of dynamic textures. That is, given a video sequence of a single dynamic texture, we want to identify which class (e.g., water, fire, etc.) the video sequence belongs to. Most of the existing dynamic texture categorization methods model the video sequence (or a manually selected image region) as the output of an LDS. Then, a distance or a kernel between the model parameters of two dynamical systems is defined.

Once such a distance or kernel has been defined, classifiers such as Nearest Neighbors (NNs) or Support Vector Machines (SVMs) can be used to categorize a query video sequence based on the training data. Among these methods, Saisan et al. used distances based on the principal angles between the observability subspaces associated with the LDSs. Vishwanathan et al. [21] used Binet-Cauchy kernels to compare the parameters of two LDSs. Further, Chan and Vasconcelos used both the KL divergence and the Martin distance as a metric between dynamical systems. Finally, Woolfe and Fitzgibbon used the family of Chernoff distances and distances between cepstrum coefficients as a metric between LDSs. Other types of approaches for dynamic texture categoriza-tion, such as Fujita

and Nayar, divide the video sequences into blocks and compare the trajectories of the states in order to perform the inference. Alternatively, Vidal and Favaro extended boosting to LDSs by using dynamical systems as weak classifiers.

3. DYNAMIC TEXTURE MODEL

Given *F* frames of a video or a spatiotemporal patch of *p* pixels,

$\{I(t) \in IR^p\}_{t=1}^{F}$, we mode $\{I(t) \in IR^p\}_{t=1}^{F}$ l the pixel intensities of each frame, *I(t),* as the output of an LDS.i.e.,

$$z(t+1)=Az(t)+Bv(t), \quad (1)$$

$$I(t) = C^0 + Cz(t) + w(t), \quad (2)$$

Where $z(t) \in IR^n$ is the $z(t) \in IR^n$ hidden state at time *t*, $A \in IR^{n \times n}$ models $A \in IR^{n \times n}$ the dynamics of the hidden state, $C \in IR^{p \times n}$ maps the $C \in IR^{p \times n}$ hidden state to the output of the system $C^0 \in IR^p$ is the $C^0 \in IR^p$ mean of the video sequence, and $w(t) \sim N(0, R)$ and $Bv(t) \sim N(0, Q)$ are the measurement and process noise, respectively. As is often the case, the transformation $Q = BB^T$ is used to incorporate the *B* matrix in the noise covariance and Bv_t is replaced by v'_t. the dimension of the hidden state, *n*, is the order of the system and *p* is the number of pixels in one frame of the sequence or patch.

Representing Videos Using the Codebook

Once the code words are available, each video sequence needs to be represented using this vocabulary. This is done by using a histogramh¼½h1;h2;...;hK€ > 2IR K . There are several choices for such a representation. In what follows, we will describe a few approaches that we will use tore present the video sequences [Figure 1]. Let us assume that codeword *k* occurs *ik* times in the *ith* video sequence. Let *N* be the total number of video sequences and *Nk* be the number of video sequences in which codeword *k* occurs at least once. The simplest representation is called the Term Frequency (*TF*) and is defined as Each of these two methods has its own advantages. The TF approach is the simplest of the approaches outlined above. Here, the focus is solely on the distribution of the codewords in a test video. The TF-IDF, on the other hand, discounts features that are common to all classes of video sequences and focuses on the ones that are unique to a particular class.

Figure 1: Sample snapshots from the UCLAB database, which is a reorganized version of the UCLA50 dynamic texture database. Each image represents a sample frame from a different video sequence in the database.

4. THEHEM ALGORITHM FORDYNAMICTEXTURES

The hierarchical expectation-maximization (HEM) algorithm was proposed in [22] to reduce a Gaussian mixture model with a large number of components into a representative GMM with fewer components. In this section, we derive the HEM algorithm when the mixture components are dynamic textures.

5. APPLICATIONS AND EXPERIMENTS

The applications exploit several desirable properties of HEM to obtain promising results. First, given a set of input DTs, HEM estimates a novel set of fewer DTs that represents the input in a manner that is consistent with the underlying generative probabilistic models by maximizing the log-likelihood of "virtual" samples generated from the input DTs. In figure 2. video texture examples. (a) video with two textures. (b) Ground-truth labels. As a result, the clusters formed by HEM are also consistent with the probabilistic framework because HEM is based on maximum-likelihood principles; it drives model estimation toward similar optimal parameter values as performing maximum-likelihood estimation on the full dataset. However, the computer memory requirements are significantly less since we no longer have to store the entire dataset during parameter estimation. In addition, the intermediate models are estimated independently of each other, so the task can be easily parallelized.

(a)

(b) river-far, escalator | grass, fire | plant-a, river-far

Figure 2: Video texture examples. (a) Video with two textures. (b) Ground truth labels.

6. CLUSTERING RESULTS

In most cases, each cluster corresponds to a single texture (e.g. grass, escalator, pond), which illustrates that HEM is capable of clustering DTs into similar motions. The Rand index for the level-2 clustering using HEM is 0.973 (for comparison, clustering histograms-of-oriented-optical-flow using K-means yields a Rand index of 0.958). One error is seen in the HEM cluster with both the river and river-far textures, which is reason-able considering that the river-far texture contains both near and far perspectives of water as given in figure 3. Moving up to the third level of the hierarchy, HEM forms two large clusters containing the plant textures (plant-i, plant-a, grass) and water textures (river-far, river, sea-far). Finally, in the fourth level, the video textures are grouped together according to broad categories: plants (grass, plant-a, plant-i), water (pond, river-far, river, sea-far), and rising textures (fire, jellyfish, and steam). These results illustrate that HEM for DT is capable of extracting meaningful clusters in a hierarchical manner.

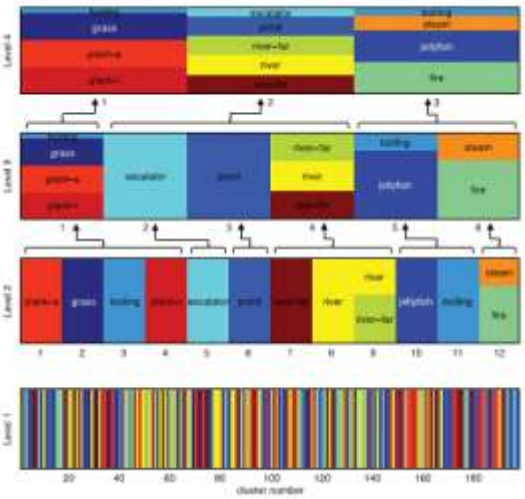

Figure 3: Hierarchical clustering of video textures: The arrows and brackets show the cluster membership from the preceding level (the groupings between Levels 1 and 2 are omitted for clarity).

7. EXPERIMENTAL SETUP

For the annotation experiment we use the DynTex dataset[23], which consists of over 650 videos, mostly in everyday surroundings. Ground truth annotation information is present for 385 sequences (called the "golden set"),

based on a detailed analysis of the physical processes underlying the dynamic textures. We select the 35 most frequent tags in DynTex for annotation comprising of 337 sequences. The tags are also grouped into two categories: 1) process tags, which describe the physical texture process (e.g., sea, field, and tree) and are mainly based on the appearance; 2) structural tags, which describe only the motion characteristics (e.g., turbulent and oscillating) and are largely independent of appearance. Note that videos with a particular structural tag can have a wide range of appearances since the tag only applies to underlying motion. Each video has an average of 2.34 tags. 1 Figure 4 shows an example of each tag alongside the number of sequences in the dataset.

Figure 4: List of tags with example thumbnails and video count for the DynText dataset. "Structural" tags are in bold.

8. CONCLUSION

In this paper, HEM approach has been adopted for categorizing dynamic textures. By modeling video with the distributions of local dynamical models extracted from it, we showed that we are able to better handle the variations in view-point and scale in the training and test data as compared to modeling the entire video sequence with a single global model. We extensively compared our algorithm with standard Bag of Features approaches using a variety of different features for categorizing video sequences as well as the original single LDS approach. Our experimental results showed that our approach produces better results across different parameter choices and empirically established the superior performance of our proposed approach. By moving from the traditional single model approach to multiple models, we can see that the performance increases. One way to further improve this model is to combine both local and global methods. Our recent approach showed that we achieve a better categorization performance by combining both local and global models.

REFERENCES:

[1] Mofakharul Islam, John Yearwood and Peter Vamplew "Unsupervised Color Textured Image Segmentation Using Cluster Ensembles and MRF Model", Advances in Computer and Information Sciences and Engineering, 323–328. © Springer Science+Business Media B.V. 2008.

[2] C.S. Wallace, and D..L. Dow, "MML clustering of multi-state, poisson, von mises circular and gaussian distribution", Statistics and Computing,Vol.10(1), Jan. 2000, pp.73-83.

[3] R. Siddheswar and R.H. Turi, "Determination of Number of Clusters in k-means Clustering and application in Color Image Segmentation", In Proceedings of the 4th Intl. Conf. on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Vol. Calcutta, India, 1999 pages: 137-143.

[4] Wu Yiming, Yang Xiangyu, and Chan Kap Luk, "Unsupervised Color Image Segmentation based on Gaussian Mixture Model", In Proceedings of the 2003 Joint Conf. of the 4th Intl. Conf. on Information Communications and Signal Processing, Vol. 1(15-18 Dec. 2003),

pages: 541-544.

[5] R. H. Turi, "Clustering-Based Color Image Segmentation", PhD Thesis, Monash University, Australia, 2001.

[6] A. Ravichandran, R. Chaudhry, and R. Vidal, "View-Invariant Dynamic Texture Recognition Using a Bag of Dynamical Systems," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[7] B. Ghanem and N. Ahuja, "Phase Based Modelling of Dynamic Textures," Proc. IEEE Int'l Conf. Computer Vision,2007.

[8] F. Woolfe and A. Fitzgibbon, "Shift-Invariant Dynamic Texture Recognition,"Proc. Ninth European Conf. Computer Vision,2006.

[9] H. Cetingul and R. Vidal, "Intrinsic Mean Shift for Clustering on Stiefel and Grassmann Manifolds," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009.

[10] A. Goh and R. Vidal, "Clustering and Dimensionality Reduction on Riemannian Manifolds, "Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.

[11] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, Mar. 2007.

[12] N. Vasconcelos and A. Lippman, "Learning Mixture Hierarchies," Proc. Neural Information Processing Systems Conf.,1998

[13] N. Vasconcelos, "Image Indexing with Mixture Hierarchies,"Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2001.

[14] G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, Mar. 2007

[15] Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on conference 13-18 June 2010

[16] Pattern Analysis and Machine Intelligence, IEEE Transactions on, IEEE Computer Society (Volume:30, Issue: 5) May 2008

[17] Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Volume:1), Conference: 17-21 Oct. 2005

[18] A. Scho¨dl, R. Szeliski, D.H. Salesin, and I. Essa, "Video Textures," Proc. ACM Siggraph, pp. 489-498, 2000.

[19] M. Szummer and R.W. Picard, "Temporal Texture Modeling," Proc. IEEE Int'l Conf. Image Processing, vol. 3, pp. 823-826, 1996

[20] S. Vishwanathan, A. Smola, and R. Vidal, "Binet-Cauchy Kernels on Dynamical Systems and Its Application to the Analysis of Dynamic Scenes,"Int'l J. Computer Vision, vol. 73, no. 1, pp. 95-119, 2007.

[21] A. Chan and N. Vasconcelos, "Classifying Video with KernelDynamic Textures,"Proc. IEEE Conf. Computer Vision and Pattern Recognition,pp. 1-6, 2007.

[22] N. Vasconcelos and A. Lippman, "Learning Mixture Hierarchies," Proc. Neural Information Processing Systems Conf.,1998

[23] P.V. Overschee and B.D. Moor, "N4SID : Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems," Automatical vol. 30, pp. 75-93, 1994.