Combustion and Flame 191 (2018) 226–238

Contents lists available at ScienceDirect

Combustion and Flame

journal homepage: www.elsevier.com/locate/combustflame

A new post-processing technique for analyzing high-dimensional combustion data

Ehsan Fooladgar*, Christophe Duwig

Linné FLOW Centre, KTH Mechanics, Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 17 July 2017
Revised 5 January 2018
Accepted 7 January 2018

Keywords:
Combustion data analysis
Complex chemistry post-processing
Stochastic embedding
Machine learning

ABSTRACT

This paper introduces a novel post-processing technique for analyzing high dimensional combustion data. In this technique, t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to reduce the dimensionality of the combustion data with no prior knowledge while preserving the similarity of the original data. Multidimensional combustion datasets are from premixed and non-premixed laminar flame simulations and measurements of a series of well documented piloted flames with inhomogeneous inlets. The resulting reduced manifold is visualized by scatter plots to reveal the global and local structure of the data (manual labeling). Unsupervised clustering algorithms are then utilized for post-processing the t-SNE manifold in order to develop an automatic labeling process.

1. Introduction

With increasing computational power, direct numerical and large eddy simulation (DNS and LES) of reacting flows with complex chemistry are becoming common, e.g. [1–3]. The resulting data which may occupy hundreds of gigabytes of storage, consists of millions to billions of points each of which is described by tens to hundreds of chemical species. To explore and analyze this large, high-dimensional data, conventional visualization techniques such as scatter plots, histograms and pairs plots are limited. Human visual perception is well tuned to identify patterns and trends in graphs with one or a few data variables at a time, calling for new automated identification tools.

One way to tackle the issue is employing computers to detect the main features of the data prior to visualization, for example by making use of recent dimension reduction methods originally developed for machine learning and computer vision. In these methods, the high-dimensional data set is converted to a two or three dimensional embedding of the data in which nearby and distant points represent similar and dissimilar points of the original data, respectively. The low-dimensional data can readily be visualized in a scatter plot or it can be used as the input for more sophisticated visualizations and/or analysis.

A popular linear dimensionality reduction method is principal component analysis (PCA). PCA projects data to a lower dimensional space with orthogonal components representing maximum variance. PCA is employed in combustion modeling [4] and identification of low-dimensional manifold in turbulent flames [5]. Moreover, its potential as a novel general approach for the analysis of reacting flows has been recently shown [6,7]. However, PCA has known limitations; namely: its modes are linear combinations of the samples, it relies on orthogonal transformations and it is sensitive to outliers. Altogether it encourages exploring more robust methods for analyzing the intrinsically nonlinear combustion data.

One of the most promising nonlinear dimension reduction methods is t-Distributed Stochastic Neighbor Embedding (t-SNE) which has been recently introduced by Van Der Maaten and Hinton [8]. t-SNE has been employed in many fields including computer security [9], music analysis [10], computer aided diagnosis [11], bioinformatics [12] and electroencephalography [13]. In this paper, we explore the first application of the tree-based t-SNE [14], developed for large-scale data, on the datasets generated from simulation and measurement of combustion phenomena. Developing an automated t-SNE-based tool for post-processing high dimensional combustion data is the other objective of this study.

The remainder of this paper is outlined as follows. Section 2 describes the t-SNE algorithm and its strengths and weaknesses. In Section 3, the performance of t-SNE is evaluated for visualization of laminar and turbulent datasets. And finally, the algorithm used for post-processing the output of t-SNE is presented in Section 4.

2. t-SNE

t-SNE is a nonparametric nonlinear dimensionality reduction method aiming to embed high-dimensional data

* Corresponding author.
E-mail addresses: efoo@mech.kth.se, ehsan.fooladgar@connect.polyu.hk (E. Fooladgar), duwig@mech.kth.se (C. Duwig).

https://doi.org/10.1016/j.combustflame.2018.01.014
0010-2180/

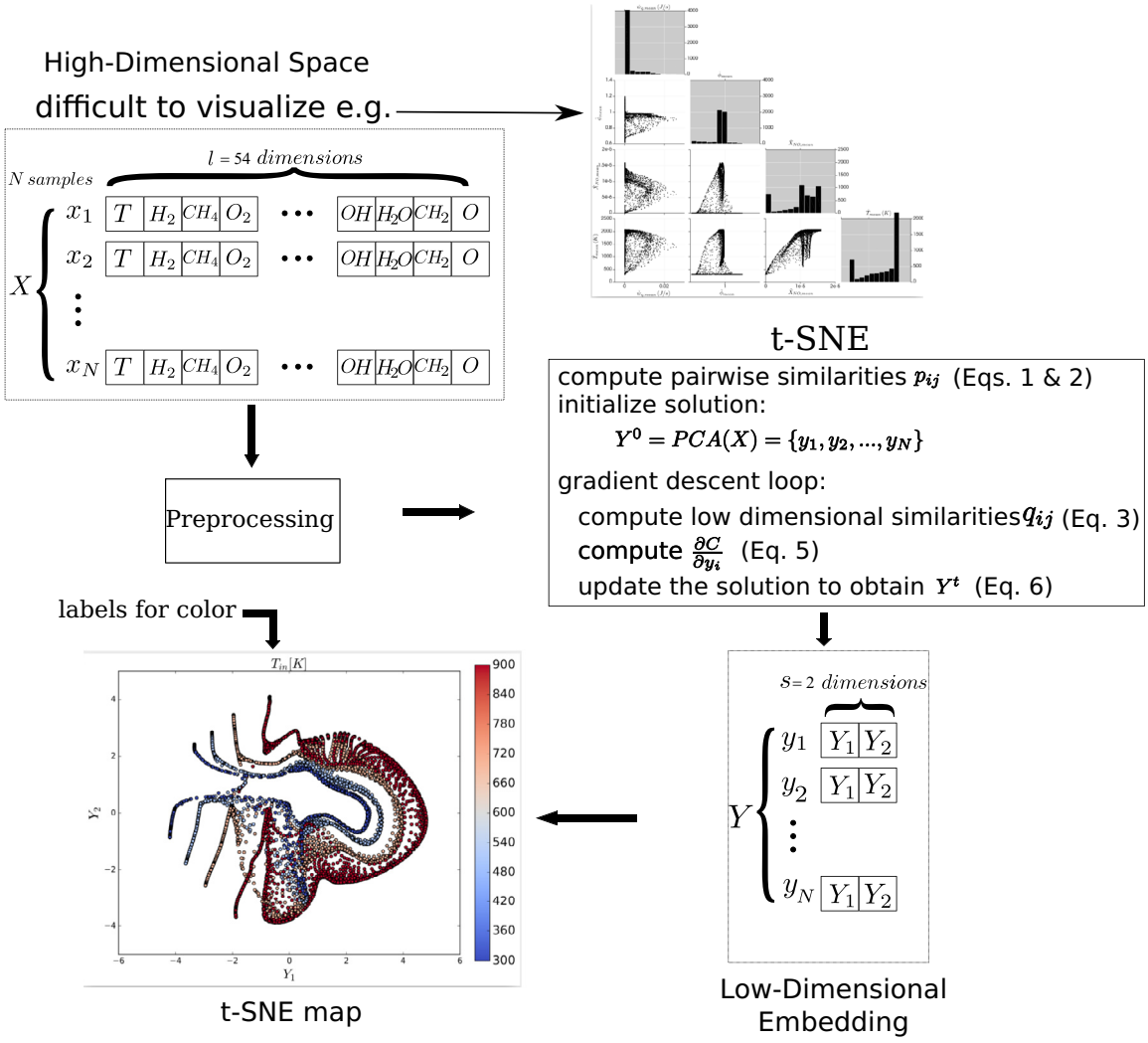

Fig. 1. A typical pipeline of combustion data visualization using t-SNE.

$X = \{x_1, x_2, \ldots, x_N\}$ with $x_i \in \mathbb{R}^l$ into a s-dimensions map $Y = \{y_1, y_2, \ldots, y_N\}$ ($y_i \in \mathbb{R}^s$) where s has a typical value of 2 or 3 and is normally much smaller than l (see Fig. 1).

The t-SNE algorithm starts by converting distances between original data points into Gaussian joint probabilities p_{ij} measuring the pairwise similarity between x_i and x_j using

$$p_{j|i} = \frac{\exp\left(-d\left(x_i, x_j\right)^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-d\left(x_i, x_k\right)^2 / 2\sigma_i^2\right)}, \qquad p_{i|i} = 0 \tag{1}$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}. \tag{2}$$

Here $d(x_i, x_j)$ is a function that returns a distance between a pair of data points and σ_i is the variance of the Gaussian centered on x_i. σ_i is determined by setting the perplexity of the conditional distribution equal to a predefined perplexity u. Therefore, σ_i is adapted to the density of the data.

In the low-dimensional map Y, the similarities q_{ij} between two points y_i and y_j are measured using a normalized Student-t kernel with one-degree of freedom:

$$q_{ij} = \frac{\left(1 + \| y_i - y_j \|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \| y_k - y_l \|^2\right)^{-1}}, \qquad q_{ii} = 0. \tag{3}$$

The heavy-tailed Student-t kernel allows dissimilar input points (i.e. points with large distance in the high-dimensional space) to be modeled by low-dimensional counterparts with a much larger distance in the map. This makes more space for the similar points, with small pairwise distances, to be modeled accurately in the low-dimensional embedding, leading to preserving the local data structure in the map.

t-SNE determines the locations of the points y_i in the map by minimizing the Kullback–Leibler divergence between the joint distributions P and Q:

$$C(Y) = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4}$$

The minimization of the objective function $C(Y)$ with respect to y_i is performed using gradient descent:

$$\frac{\partial C}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \| y_i - y_j \|^2\right)^{-1}, \tag{5}$$

by computing which, the solution at iteration t, Y^t, is obtained using

$$Y^t = Y^{t-1} + \eta \frac{\partial C}{\partial y_i} + \mu(t)\left(Y^{t-1} - Y^{t-2}\right), \tag{6}$$