A Block-separable Parallel Implementation for the Weighted Distribution Matching Similarity Measure

Mauricio Orozco-Alzate¹, Eduardo-José Villegas-Jaramillo¹, and Ana-Lorena Uribe-Hurtado¹*

Universidad Nacional de Colombia - Sede Manizales, Facultad de Administración, Departamento de Informática y Computación, Grupo de Ambientes Inteligentes Adaptativos - GAIA, km 7 vía al Magdalena, Manizales 170003, Colombia
{morozcoa, ejvillegasj, alhurtadou}@unal.edu.co

Abstract. Automatic pattern recognition is often based on similarity measures between objects which are, sometimes, represented as high-dimensional feature vectors —for instance raw digital signals or high-resolution spectrograms. Depending on the application, when feature vectors turn extremely long, computations of the similarity measure might become impractical or even prohibitive. Fortunately, multi-core computer architectures are widely available nowadays and can be efficiently exploited to speed-up computations of similarity measures. In this paper, a block-separable version of the so-called Weighted Distribution Matching similarity measure is presented. This measure was recently proposed but has not been analyzed until now for a parallel implementation. Our analysis shows that this similarity measure can be easily decomposed into subproblems such that its parallel implementation provides a significant acceleration in comparison with its corresponding serial version. Both implementations are presented as Python programs for the sake of readability of the codes and reproducibility of the experiments.

Keywords: Similarity measure, parallel algorithm, weighted distribution matching, block separability, multi-core implementation

1 Introduction

Measuring the degree of similarity among objects is an essential issue in several pattern recognition fields such as information retrieval, clustering and classification. Objects in these fields are typically observed as digital signals (data sequences, images or videos) and afterwards represented as feature vectors which, according to their nature, can be compared by a wealth of similarity measures that has been proposed in the literature [2, 8]. Such similarity measures must be fast, either by definition or implementation, as well as accurate to satisfactorily capture the often task-dependent concept of resemblance between objects.

* Estudiante del Doctorado en Ingeniería, Industria y Organizaciones - Universidad Nacional de Colombia - Sede Manizales.
The simplest and most straightforward feature representation consists in using the signal samples themselves, for instance the pixels in the image ordered as a long feature vector. A more elaborate but still simple representation is the computation of a histogram, which provides an estimation of the data distribution by partitioning the range of admissible sample values of the signal in bins and counting the number of samples that fall in each of them. A properly normalized version of the histogram can be considered an estimate of the probability mass/density function and, therefore, it is very useful to characterize and discriminate among observations.

Pixel representations and histograms tend to generate high-dimensional feature vectors. Consider, for instance, the following two exemplar cases: first, images from high-definition cameras such as the ones used in dermatopathology with typical resolutions of $920 \times 1080 = 993,600$ pixels and, second, full histograms for RGB color images whose three bands are encoded with one byte each and, therefore, $2^8 \times 2^8 \times 2^8 = 16,777,216$ different colors can be generated [3]. Even though high-dimensional feature representations exhibit known disadvantages such as non-convex subspaces and potential loss of information of the spatial connectivity, they are still valuable representations to be used, specially when sufficient training objects are available or in case that no good and more advanced features can be found [6].

Amongst the plethora of available similarity measures, the so-called *Weighted Distribution Matching* (WDM) [5] stands out due to its fulfillment of the above-mentioned desired properties of speed and accuracy as well as because it takes into account the structure of the feature vector, namely the relationship between consecutive entries. In spite that the WDM similarity measure has a low computational cost, a detailed analysis of its mathematical structure—aimed at revealing possibilities for decomposing the algorithm into subproblems—has not been undertaken yet. Such an analysis will serve to formulate a parallel implementation of the WDM similarity measure that takes advantage of the nowadays omnipresent multi-core computer architectures with the aim of speeding up its computation when comparing high-dimensional feature vectors.

In this paper, we carry out an analysis of the mathematical formulation of the WDM similarity measure, particularly looking for block separability [4] such that partial results can be computed independently on different computer cores. For the sake of reproducibility and readability, Python implementations are presented. They, of course, could be speeded up further by translating the implementations to a compiled language if the user requires an optimized solution to be deployed instead of a flexible multi-core implementation for simulation purposes. The remaining part of the paper is organized as follows. The original sequential algorithm of the WDM similarity measure is presented in Section 2. The proposed block-separable version of it is described in Section 3. Experiments showing the performance of both sequential and parallel implementations of the WDM similarity measure, for comparing high-dimensional feature vectors as those resulting from high-resolution images and full color histograms, are presented in Section 4. Finally, our conclusions are drawn in Section 5.