1. Methods overview
   1. Metrics Distances
      1. DTW
      2. ERP
      3. EMD
   2. Features
      1. Shape Context
      2. Multi Angular Descriptor
   3. Fast Retrieval data structure
      1. Kdtree
      2. LSH
   4. Dimensionality Reduction
      1. PCA
      2. LDA
   5. Learning Models
      1. SVM
      2. WaldBoosting

## Dynamic Time Warping

***Dynamic time warping*** (DTW) is an algorithm for measuring similarity between two time serieses. The method is to find an optimal alignment between two given sequences. Intuitively, the sequences are warped in a non-linear fashion to match each other. DTW solves the discrepancy between intuition and calculated distance by achieving the optimal alignment between sample points of two times sequences. Restrictions can be imposed on the matching technique to improve the matching results and the algorithm’s time complexity. The only restriction placed on the data sequences is that they should be sampled at equidistant points in time which can be easily achieved by re-sampling. A naïve approach to calculate the distance between two time serieses can be to resample one of the serieses and to compute the distance sample by sample, the problem with this technique is that it does not yield intuitive results as it might match samples that do not correspond well. See figure 1.

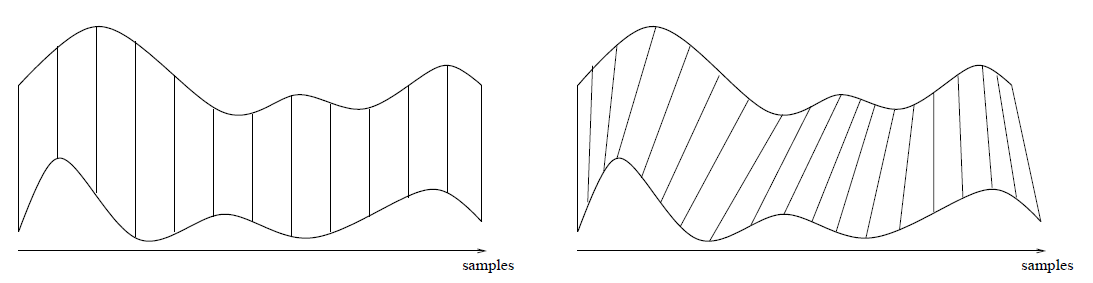


Figure 1 – The right scheme shows sample-by-sample naïve alignment after resampling and in the left scheme the alignment was performed using DTW

Formally, given two time serieses, and , DTW yields an optimal warping path, where  for which satisfies the two following:

1. **Start and End point constrain:** and.
2. **Time preservation of the points:** for each two points in W,,  and the following holds: and .
3. **Local continuity constrain:**

The Distance of the warping path is defined as follows:

|  |  |  |
| --- | --- | --- |
|  |  | Equation 1 |

 is the distance of the two data point indexes (one from X and the other from Y) in the kth element of the warping path.

The warping path which has the minimal  associated with alignment is the optimal warping path which we have previously denoted by.

Using dynamic programing approach, DTW yields the optimal warping path by constructing the **accumulated distance matrix.** We will denote this matrix by the letter, where is the minimum distance warping path that can be constructed from the two time serieses and. Wherefore the value in  will contain the minimum-distance warping path between  and .

The accumulated distance matrix is calculated as follows:

|  |  |  |
| --- | --- | --- |
|  |  | Equation 2 |
|  |  |  |
|  |  | Equation 3 |

Once matrix D is calculated the optimal warping path  is retrieved by backtracking the matrix  from the point  to the point following the greedy strategy of looking for the direction from which the current distance is taken.

It is easy to note that the time and space complexity of DTW is. [7] [8] [9]

### DTW Speedup

One drawback of DTW is its quadratic time and space complexity, thus, many speedup methods have evolved. These methods can be fall in three categories:

1. Constrains: Limiting the amount of calculated cells in the cost matrix.
2. Data Abstraction: Running the DTW on an abstract representation of the data.
3. Lower Bounding: Reducing the times DTW need to run when clustering and classifying time serieses.

Constrains: The most popular two constrains on DTW are the Sakoe-Chuba Band and he Itakura parallelogram, which are shown in figure 2. The grayed out area is the cells of the cost matrix that are filled by the DTW algorithm for each constrain. The warping path is looked for in the constrain window. The width of the window is specified by a parameter. By using such constrains the speedup factor is a constant and the DTW complexity is still (where M and N are the time serieses lengths). Note that if the warping path is does not reside in the constrain window, it will not be found by DTW, thus, such method is usually used when the warping path is expected to be in the constrain window.

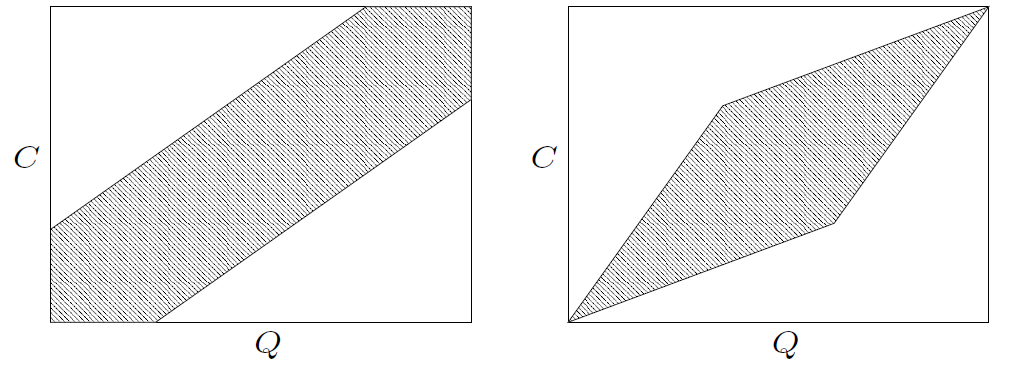


Figure 2 – Cost matrix constrains: Sukoe-Chuba Band (left) and the Itakura Parallelogram (right).

Data Abstraction: Speedup using data abstraction is performed by running DTW on a reduced presentation of the data thus reducing the cell numbers that need to be computed. The warp path is calculated on the reduced resolution matrix and mapped back to the original (full) cost matrix.

Lower bounding: Searching the most similar time series in the database given a template time series can be done more efficiently using lower bound functions than using DTW to compare the template to every series in the database. A lower-bounding is cheap and approximate. However, it underestimates the actual cost determined by DTW. It is used to avoid comparing serieses by DTW when the lower-bounding estimate indicates that the time series is worse match than the current best match.

FastDTW, proposed by Stan Salvador and Philip Chan, approximate DTW in a linear time using multilevel approach that recursively projects a warping path from the coarser resolution to the current resolution and refines it. This approach is an order of magnitude faster than DTW, and also compliments existing indexing methods that speedup time series similarity search and classification. [9]

Ratanamahatana and Kough have shown the following fact about DTW in their work in [10]: Comparing sequences of different lengths and reinterpolating them to equal length produce no statistically significant difference in accuracy or precision/recall.

Although in our work we do make reinterpolating of sequeces to compare equilenth sequences.

## Earth Movers Distance

Earth movers distance (EMD) is a measure of the dissimilarity between two histograms. Descriptively, if the histograms are interpreted as two different ways of piling up a certain amount of dirt, the EMD is the minimal cost of turning one pile to other, where the cost is assumed to be the amount of dirt moved times the distance by which it moved.

EMD has been experimentally verified to capture well the perceptual notion of a difference between images. [11]

[Add a picture]

Computing EMD is based on a solution to the well-known transportation problem. It can be computed as the minimal value of a linear program.

[EMD formal definition]

EMD naturally extends the notion of a distance between single elements to that of a distance between sets, or distributions, of elements. In addition, it’s a true metric if the signatures are equal. [12] A major hurdle to using EMD is its computational complexity (for an N-bin histogram).Various approximation algorithms have been proposed to speed up the computation of EMD.

A histogram can be presented by its signature. The signature of a histogram is a set of clusters where each cluster is represented by its mean (or mode), and by the fraction of the histogram that belongs to that cluster.

### EMD Embedding

The Idea of the Embedding is to compute and concatenate several weighted histograms of decreasing resolutions for a given point set. The embedding of EMD given in [11] provides a way to map weighted point sets A and B from the metric space into the normed space , such that the  distance between the resulting embedded vectors is comparable to the EMD distance between A and B. The motivation of doing so is that the working with normed space is desirable to enable fast approximate Nearest Neighbors (NN) search techniques such as LSH.

[Embedding formal definition]

### Wavelet EMD

A work conducted by Shirdhonkar and Jacobs in [13] has proposed a method for approximating the EMD between two histograms using a new on the weighted wavelet coefficients of the difference histogram. They have proven the ratio of EMD to wavelet EMD is bounded by constants. The wavelet EMD metric can be computed in time complexity.

[Embedding formal definition]

## Features

The selection of valuable features is crucial in pattern recognition.

## Shape context

S. Belongie et al. have defined in [14], a point matching approach named Shape Context. The Shape context is a shape matching approach that intended to be a way of describing shapes that allows for measuring shape similarity and the recovering of point correspondences. This approach is based on the following descriptor: Pick n points on the shape’s contour, for each point  on the shape, consider the  other points and calculate the coarse histogram of the relative coordinates such that is defined to be the shape context of. The bins are normally taken to be uniform log-polar space. This distribution over relative positions is robust and compact, yet highly discriminative descriptor.

To match two points, from the first shape and from the other shape, define  denote the cost of matching, these two points, use the  test statistic:

|  |  |
| --- | --- |
|  | Equation 4 |

.

To perform a one-to-one matching that matches each point  on shape 1 and  on shape 2 that minimizes the total cost of matching: is needed.

This can be calculated in  time using the Hungarian method.

[Add a picture]

## Multi Angular Descriptor

The Multi Angular Descriptor (MAD) is a shape recognition method described in [15], which captures the angular view to multi resolution rings in different heights. Given an Image I of a connected component, calculate the centroid C and the diameter D of the image I. draw a set of rings centered by C with different radius values which are derived from the diameter D. draw k points on each ring taken with uniform distance from each other. Each point in each ring serves as an upper view point watching each contour point in the shape.

## Fast retrieval data structures

Now we will give an overview on 2 methods that tries to answer the second question.

### K dimensional tree

## Support Vector machine

SVM (Support Vector Machine) is a powerful tool for classification developed by Vapnik in 1992. It has originally been proposed for two-class classiﬁcation; however, it can be easily extended to solve the problem of multiclass classification. In this work we use one-to-one/ one-to-many method to do so. SVM is widely used in object detection & recognition, content-based image retrieval, text recognition, biometrics, speech recognition, etc. Without Losing generality, we consider the 2 class classification problem. We give a brief introduction for the SVMs basics. SVM paradigm has a nice geometrical interpretation of discriminating one class from the other by a separating hyperplane with maximum margin. SVM can be altered to perform nonlinear classification using what is called the kernel trick by implicitly mapping the inputs into high-dimensional feature spaces. Given a training sample set  where a d-dimensional sample,  is the class label and N is the size of the training set.

Support vector machine first map the data from the sample from the input space to a very high dimensional Hilbet Space.. The mapping  is implemented implicitly by a kernel function  that satisfies Mercer’s conditions. Kernels are functions that give the inner product of each pair of vectors in the feature space, such that. Then in the high dimensional feature space H, we try to find the hyperplane that maximizes the margin between the two classes, and minimizing the number of training error.

The decision function is as follows:

|  |  |
| --- | --- |
|  | Equation 4 |

Where

|  |  |
| --- | --- |
| And | Equation 5 |

Training as SVM is to find  that minimizes the following quadratic cost:

|  |  |
| --- | --- |
|  | Equation 6 |

Where  is a parameter chosen by the user, a large  corresponds to a higher penalty allocated to the training errors. This optimization problem can be solved using quadratic programming. [16]

Many implementation of kernels have been proposed, one popular example is the Gaussian Kernel:

|  |  |
| --- | --- |
|  | Equation 7 |
|  |  |

[Add a picture]

## Gausian dynamic time warping using DTW as a metric

Claus Bahlmann et al. in [17] have proposed combining DTW and SVMs by establishing a new kernel which they have named Gaussian DTW (GDTW) kernel. They have achieved, by incorporating DTW in the Gaussian kernel function, a method to handle online handwriting input which is a variable-sized sequential data. Clearly, the assimilation of the DTW in the Gaussian kernel will result the following kernel function

|  |  |
| --- | --- |
|  | Equation 8 |
|  |  |

We should note that the DTW is not a metric, as the triangle inequality do not hold in many cases, thus the GDTW may miss some important properties as satisfying the Mercer’s condition. This would imply that the optimization problem in 6 may not convergent to the global minimum.

### Principle Component Analysis

PCA was invented in 1901 by Karl Pearson. It is a linear technique for dimensionality reduction, which means that it performs dimensionality reduction by embedding the data into a linear subspace of lower dimensionality. Although there exist various techniques to do so, PCA is by far the most popular (unsupervised) linear technique. PCA is an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (names the first principal component), the second greatest variance on the second coordinate, and so on. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data.

The full set of principal components is as large as the original set of variables. But taking the first few principal components will preserve most of the information in the data, and reduces the data dimensions.

### Linear Discrimination Analysis

PCA is an unsupervised technique and as such does not include label information of the data. The following example demonstrates the problem drawback: Imagine 2 cigar like clusters in 2 dimensions, one cigar has y = 1 and the other y = -1. The cigars are positioned in parallel and very closely together, such that the variance in the total data-set, ignoring the labels, is in the direction of the cigars. For classiﬁcation, this would be a terrible projection, because all labels get evenly mixed and we destroy the useful information. A much more useful projection is orthogonal to the cigars, i.e. in the direction of least overall variance, which would perfectly separate the data-cases.

LDA is closely related to PCA in that they both look for linear combination of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. In this method, variability among the feature vectors of the same class is minimised and the variability among the feature vectors of different classes is maximised.

The LDA performs dimensionality reduction while preserving as much of the class discriminatory information as possible. Without going into the math, in order to find a good projection vector, we need to define a measure of separation between the projections. The solution proposed by Fisher is to maximize a function that represents the difference between the means, normalized by a measure of the within-class scatter.

Although, LDA assumes that the distribution of samples in each class is Gaussian and that we cannot prove that the handwritten letters are distributes in a Gaussian manner, we selected LDA as

[put the picture that demonstrates the problem with PCA]

[Should I Add mathematical explanation?]