# Efficient Word Image Retrieval using Fast DTW Distance

G. Nagendar and C.V. Jawahar
Center for Visual Information Technology, IIIT Hyderabad, India
Email: nagendar.g@research.iiit.ac.in, jawahar@iiit.ac.in

***Abstract*—Dynamic time warping (DTW) is a popular distance measure used for recognition free document image retrieval. However, it has quadratic complexity and hence is computationally expensive for large scale word image retrieval. In this paper, we use a fast approximation to the DTW distance, which makes word retrieval efficient. For a pair of sequences, to compute their DTW distance, we need to find the optimal alignment from all the possible alignments. This is a computationally expensive operation. In this work, we learn a small set of global principal alignments from the training data and avoid the computation of alignments for query images. Thus, our proposed approximation is significantly faster compared to DTW distance, and gives 40 times speed up. We approximate the DTW distance as a sum of multiple weighted Eulidean distances which are known to be amenable to indexing and efficient retrieval. We show the speed up of proposed approximation on George Washington collection and multi-language datasets containing words from English and two Indian languages.**

## I. Introduction

Word image retrieval from a large corpus of document images is a challenging problem. The problem has been looked in two settings: recognition based [7, 14] and recognition free [9, 16]. Recognition free based approach have gained interest in recent years. It has two primary dimensions, (i) represent word images, and (ii) compare word image representations. Word spotting [9] is a promising method for recognition free retrieval. In this method, word images are represented using different features, and the features are compared with the help of appropriate distance measure. Word spotting has the advantage that it does not require prior learning due to its appearance based matching. These techniques have been popularly used in document image retrieval. For example, searching documents in a collection of printed documents [3], accessing handwritten documents [9] etc. In this technique, word images are often compared using Dynamic Time Warping (DTW) [2, 13]. In general, word spotting technique with DTW works well. However, for comparing two word images, it typically takes one second [9]. This is computationally not attractive. To get faster retrieval, we need an efficient way of computing DTW distance.

Distance/Similarity measure plays an important role in word image retrieval system. It is used to compare two word representations. DTW distance popularly used for comparing word image representations [11]. This is mainly due to its ability to capture local dependencies, and handling variable length representations. DTW distance has been successfully applied in many areas like, bioinformatics [1] and word recognition [4, 8]. DTW distance works well on many classification and retrieval problems with Nearest Neighbour classifier [15]. For a pair of given two sequences, to compute their DTW distance, we need to find the optimal alignment which has the least cost from all the possible alignments. This is computationally expensive operation in DTW distance. For given two sequences of length $n$ and $m$ respectively, its computational complexity is $O(nm)$. Compared to DTW distance, Euclidean distance can be computed in linear time, however, its main limitation is that it does not capture the local dependencies. In this work, we approximate the DTW distance as a sum of multiple weighted Eulidean distances which are known to be amenable to indexing and efficient retrieval. For speed-up, DTW distance is previously approximated [6, 10] using different techniques.

For a given set of sequences, there are similarities between the top alignments (least cost alignments) of different pairs of sequences. In this work, we explore these similarities by learning a small set of the global principal alignments from the training data. To compute the global principal alignments, first, we compute the top alignments for all pairs of samples from the given training data, and then the global principal alignments are computed from these top alignments. We use 2D-PCA for computing the global principal alignments. For a given new query, instead of computing the optimal alignments, we use only these global principal alignments for computing the fast approximate DTW distance. Since we avoid the computation of alignments the proposed approximation is computationally efficient compared to naive DTW distance.

The main contribution of the paper is fast approximation of DTW distance. This is achieved by introducing global principal alignments, which avoids computation of optimal alignments for new query images. These precomputed global alignments captures all the correlations in the datasets. The proposed fast approximation of DTW distance makes DTW based document image retrieval computationally feasible. We have demonstrated utility of the proposed technique for retrieving word images from popular George Washington database and Indian language datasets. We show the superiority of proposed approximation by comparing it with naive DTW distance, Euclidean distance and metric DTW. The proposed technique is computationally efficient compared to DTW distance and metric DTW with very minor drop in retrieval performance. On multi language datasets, we show a speed up of more than 40 times compared to DTW distance and metric DTW. In summary, the performance of proposed method is as good as DTW distance and computationally performs equally as simple Euclidean based matching.

The paper is organized as follows. The next section describes the popular dynamic time warping technique. The overview of our proposed approximation technique is dis-

cussed in Section III. The experimental settings and evaluation protocols are discussed in Section IV. Section V discusses experimental evaluations of the proposed method, followed by concluding remarks in Section VI.

## II. Dynamic Time Warping

Dynamic Time Warping [12] (DTW) is used to compute the similarity/dissimilarity between two sequences. For a given two sequences, it finds the matching distance by computing the optimal alignment between the sample points. This optimal alignment has minimal cumulative distance corresponding to the aligned sample points compared to all the possible alignments. It can handle local distortions in word images. For a pair of two sequences $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$ ($x_i$s, $y_i$s are sample points) of length $n$ and $m$ respectively, an alignment $\rho$ of length $t$ is a pair of increasing $t$-tuples $(\rho_1, \rho_2)$ such that

$$1 = \rho_1(1) \leq \ldots \leq \rho_1(t) = n,$$
$$1 = \rho_2(1) \leq \ldots \leq \rho_2(t) = m$$

with unitary increments and no simultaneous repetitions, i.e. $\forall 1 \leq i \leq t-1$,

$$\rho_1(i+1) \leq \rho_1(i) + 1, \;\; \rho_2(i+1) \leq \rho_2(i) + 1$$
$$(\rho_1(i+1) - \rho_1(i)) + (\rho_2(i+1) - \rho_2(i)) \geq 1.$$

An alignment between two sequences gives a way of matching the sample points of one sequence to another. Let $S(X, Y)$ be the set of all alignments between $X$ and $Y$. The DTW distance between the sequences $X$ and $Y$ gives an optimal alignment which has minimum distance compared to all the alignments, and the corresponding distance is their DTW distance. This is given as

$$DTW(X, Y) = \min_{\rho \in S(X,Y)} \psi(X_{\rho_1}, Y_{\rho_2})$$

where, $X_{\rho_1} = (x_{\rho_1(1)}, \ldots, x_{\rho_1(|\rho|)})$ and $\psi(X_{\rho_1}, Y_{\rho_2})$ is ground distance between the points $X_{\rho_1}$ and $Y_{\rho_2}$. In general, Euclidean distance is used as the ground distance.

## III. Approximating DTW distance

In general, DTW distance has quadratic complexity in length of the sequence. In this section, we present a linear approximation to the DTW distance. For a pair of samples, DTW distance is computed using the optimal alignment from all the possible alignments. This optimal alignment captures all the linear and non-linear correlations between the given sequences. Computation of optimal alignment is the most expensive operation in finding DTW distance.

For a given set of sequences, there are similarities between the optimal alignments of different pairs of sequences. For example, if we take two different classes, the top alignments between the samples from these classes always have some similarity. Based on this idea, we compute a set of global principal alignments from the training data and use these alignments for computing the DTW distance between new test sequences. This avoids the computation of optimal alignments.

**Algorithm 1** Fast approximation of DTW distance.

**Input:** Feature vectors of word images.
**Output:** Global principal alignments.
**Step 1:** Compute the top alignments for given feature vectors.
**Step 2:** Represent the alignments using 2D matrices.
**Step 3:** Apply 2D-PCA over the alignment matrices and compute global principal alignments.

Now, the DTW distance becomes sum of the Euclidean distances over the global principal alignments. This gives a linear approximation of the DTW distance. These alignments captures all the similarities in the sequential data.

We use principal component analysis (PCA) for modeling the global principal top alignments for the the given data. PCA is a statistical procedure that uses an orthogonal transformation to convert a set of correlated observations into a set of values of linearly uncorrelated variables called principal components. The number of principal components are less than or equal to the number of original variables. This transformation computes the principal components in such a way that the first principal component has the largest possible variance and accounts for as much of the variability in the data as possible. The next principal component has next highest possible variance. The objective of our work is to compute the global principal alignments using the set of all alignments such that the computed alignments should be well enough for approximating the DTW distance between any new pair of sequences. This is similar to the concept of PCA, where global principal alignments correspond to eigenvectors. However, for word document images, it will be difficult to learn the global principal alignments using PCA due to variable size of word images, this leads to variable size feature representation (sequences). Also the corresponding length of sequences is large for rich feature representation. This creates a large dimensional covariance matrix and computing the eigenvalues and eigenvectors for this matrix will be computationally expensive. Similar to PCA, a two-dimensional principal component analysis (2D-PCA) [17] is proposed to overcome these issues. 2D-PCA works on 2D matrices rather than 1D vectors. In this work, instead of 1DPCA, we use 2D-PCA for computing the global principal alignments. To compute the global principal alignments for a given data, we first represent each alignment using a 2D matrix and then global principal alignments are computed by applying 2D-PCA over these alignment matrices.

To apply 2D-PCA, we need to represent each alignment using a 2D matrix. For a given pair of word images, its corresponding alignments and their alignments matrix representation is given in Figure 1. First, we compute feature representation of given word images, this gives feature vectors (sequences) for each word image. For a given two sequences of length $n$ and $m$, an alignment $\rho$ between them can be represented by using an $n \times m$ grid. This can also be represented using an $n \times m$ binary matrix, where the elements in this matrix are either 0 or 1. The entries through which the alignment passes are 1 and other entries are 0. This representation is shown in Figure 1 (c)-(d). The complete algorithm for proposed approximation technique is shown in Algorithm 1.

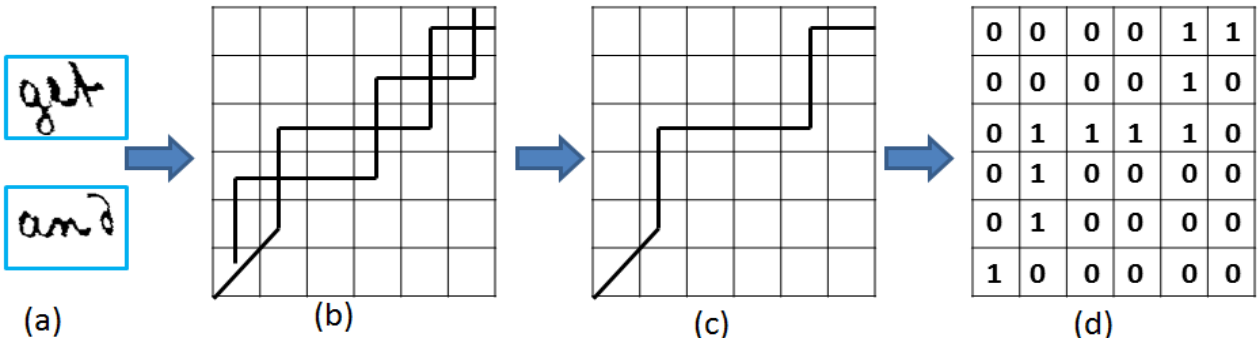


Fig. 1: Alignment matrix representation. (a) the given two images. (b) two possible alignments between the given images. (c) one possible alignment. (d) alignment matrix corresponding the alignment in (c).

## IV. Retrieval of Word Images

In this section, we discuss various components involved in our proposed approximation technique. We extract features from the image, and match them with those computed for each word in the database.

### A. Feature Extraction

In this work, we use the split profile features [9] for representing the word images. In this, we divide the image horizontally into two parts and the following features are computed. (i) vertical profile i.e the number of ink pixels in each column (ii) location of lowermost ink pixel, (ii) location of uppermost ink pixel and (iv) number of ink to background transitions. The profile features are calculated on binarized word images obtained using the Otsu thresholding algorithm.

### B. Metric DTW distance

DTW distance computes all the non-linear similarities between the given sequences. However, it is not a metric. There is a popular metric DTW distance, which commonly used in kernel setting, especially with SVM. It performs better compared to DTW distance. However, it is computationally costly. Compared to DTW distance, where it uses only optimal alignment, the metric DTW considers all the possible alignments. Thus, it performs better compared to DTW distance.

The metric DTW is defined as follows

$$\kappa_{DTW} = exp\Big( - \sum_{\rho \in S(X,Y)} \frac{1}{|\rho|} \sum_{i=1}^{|\rho|} \|x_{\rho_1(i)} - y_{\rho_2(i)}\|_2^2 \Big) \quad (1)$$

Using the proposed technique, we can also approximate this metric DTW.

### C. Precomputing the alignments

For a given dataset, we precompute the global principal alignments as follows. Since for any given two sequences there exist possibly many alignments, there will be possibly many alignment matrices. Let for a given two sequences $X$ and $Y$, denote $\rho^1, \ldots, \rho^t$ as their total number of possible alignments. Let us assume that these alignments are arranged according to their cost, i.e. the alignment $\rho^1$ has the least cost and $\rho_t$ has the maximum cost. The alignment $\rho^1$ will be the optimal alignment between $X$ and $Y$. For these alignments, we denote their corresponding alignment matrices as $A^1_{X,Y}, \ldots, A^t_{X,Y}$ and denote these total alignment matrices as follows

$$A_{X,Y} = \cup_{i=1}^{t} A^i_{X,Y}$$

where, the matrix $A^i_{X,Y}$ corresponds to the alignment $\rho_i$ and $t$ is the total number of alignments. Let us denote the total alignment matrices for a given dataset $\mathcal{X}$ as follows

$$T_{\mathcal{X}} = \bigcup_{X,Y \in \mathcal{X}} A_{X,Y}.$$

Since there exists possibly many alignments for every pair of sequences, the cardinality of $T_{\mathcal{X}}$ will be huge. Due to this, instead of computing global principal alignments over $T_{\mathcal{X}}$, we compute them using a subset of $T_{\mathcal{X}}$. This subset consists of alignment matrices corresponds to only top alignments. Since only top alignments have significant role in computing DTW distance, we consider only top few alignments. Let us define this subset as follows,

$$T^r_{\mathcal{X}} = \bigcup_{X,Y \in \mathcal{X}} (\cup_{i=1}^{r} A^i_{X,Y}).$$

Now, the set $T^r_{\mathcal{X}}$ consists of only top $r$ alignment matrices for every pair of sequences. Define the set $P_t$ as follows,

$$P_t = \cup_{X,Y \in \mathcal{X}} (\cup_{j=1}^{t} B^j_{X,Y}).$$

The set $P_t$ takes only top $t$ alignments between every pair of samples. Now the set $P_t$ contains all the best possible alignment matrices for the given dataset. We compute the global principal alignments from this set of alignment matrices. We find these alignments using 2D-PCA. If the dataset contains alignments of variable length, the set $P_t$ contains matrices of variable dimension. In this case, we cannot apply the above procedure. To overcome this, we first scale the alignments to a fixed size then alignment matrices are computed from these scaled alignments. The resulting alignment matrices will have the same dimension. We then apply 2D-PCA over this set of matrices and find the eigenvectors. These eigenvectors give the global principal alignments for the given data.

### D. Data sets and evaluation protocols

In this sub section, we discuss the datasets and the experimental settings that we follow in the experiments. We show

| Dataset | # classes | # images |
|---|---|---|
| D1 | 125 | 16145 |
| D2 | 268 | 30164 |
| D3 | 100 | 14306 |
| George Washington Database (GW) | 1471 | 4894 |

TABLE I: Details of the datasets considered in the experiments.

| | D1 | D2 | D3 | GW |
|---|---|---|---|---|
| # Samples | 16145 | 30164 | 14306 | 4894 |
| # Global Alignments | 60 | 100 | 60 | 40 |

TABLE II: Number of global principal alignments for the datasets used in the experiments. Here, the number of global alignments are based on the size of the dataset.

the results on popular George Washington (GW) database. In the experimental section, to demonstrate the utility of the proposed method across languages, we also show the results on various Indian languages. Datasets contain two different Indian languages (Hindi (D1) and Telugu (D2)) and English (D3) with significant change in structure. One of the Indian languages have a headline and the other does not have. One of them is Aryan language and the other one is Dravidian language. Details of the dataset are given in Table I. For datasets D1, D2 and D3, ground truth is created using [5].

To evaluate the quantitative performance, multiple query images were generated. The query images are selected such that, they have multiple occurrences in the database, and are mostly functional words, also they have no stop words. The performance is measured by mean Average Precision (mAP). The mAP is the mean of the area under the precision-recall curve for all the queries. For every pair of time series, we take $t = 10$, i.e. we choose top 10 alignments. For all the datasets, we take the number of global principal alignments based on their size. The number of global principal alignments for each of the dataset is given in Table II. As we compute global principal alignments using 2D-PCA, we need fixed size alignment matrices. Since, size of the images is not fixed in our datasets, we scale each word image into a fixed size. All experiments were carried out on a single core of a 2.1 GHz AMD 6172 processor with 12 Gb RAM.

## V. Results and Discussions

In the first experiment, we show the performance of proposed Fast approximation of DTW distance. We compare proposed method with naive DTW distance, Euclidean distance and metric DTW. The results over the given datasets are shown in Table III. We compare the mAP score over all the datasets. Proposed approximation is almost comparable to DTW distance over all the datasets. It performs significantly better compared to Euclidean distance. In all the four methods metric DTW performs well over all the datasets. This is mainly due to the consideration of all the alignments instead of considering only optimal alignment. For given sequences, some times only optimal alignment may not be sufficient, in this case considering all the alignment or few top alignments will enhance the performance. The slight drop in the performance

| | DTW distance | Proposed | Metric DTW | Euclidean |
|---|---|---|---|---|
| D1 | 0.9038 | 0.8927 | 0.9127 | 0.8059 |
| D2 | 0.8382 | 0.8204 | 0.8438 | 0.7123 |
| D3 | 0.8193 | 0.8072 | 0.8319 | 0.7329 |
| GW | 0.5173 | 0.5019 | 0.5361 | 0.3271 |

TABLE III: Performance of the proposed technique as compared to the DTW distance, metric DTW and Euclidean distance. Here, the mAP score is compared for all the methods.

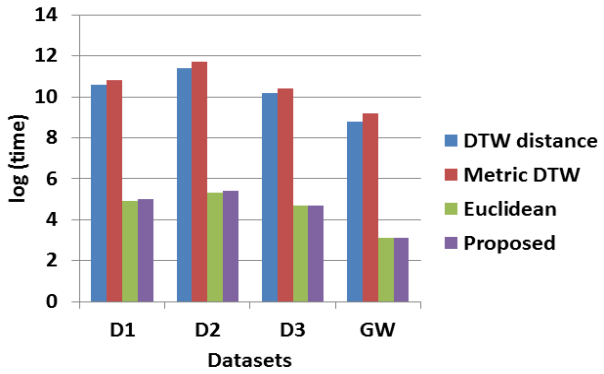


Fig. 3: Retrieval time for a given query image for all the four methods over the given datasets. Here, the retrieval time is shown over log scale. Clearly, the proposed approximation is computationally efficient over all the datasets.

of proposed approximation compared to DTW distance is due to the scaling step and the size of training data. Our proposed approximation is comparable to the DTW distance and metric DTW but computationally it is faster than these methods.

To explore the speed up of proposed approximation technique, we compare its retrieval time with naive DTW distance, Euclidean distance and metric DTW over all the datasets. The results are shown in Figure 3. Here, the retrieval time is shown over log scale. Fast approximation of DTW distance is significantly faster compared to DTW distance and metric DTW, and almost comparable to Euclidean distance. This is mainly due to the length of the feature vectors. In the Fast DTW distance, we project the data in to more than one principal directions, due to which the resulting representation has larger dimension compared to the original representation. The speed up of our proposed technique is due to the global principal alignments, which we are computing from the training data. For a given query image, in DTW distance, we need to find the optimal alignments with all the images in the database. In the proposed technique, we use precomputed global principal alignments for computing the distance. Due to this the proposed approximation is computationally efficient compared to other methods. Note that being computationally attractive our method achieve comparable performance to DTW based retrieval.

To evaluate the applicability of fast approximation of DTW distance across various languages, we show experimental results on printed Indian language datasets namely D1 and D2. In general, these languages need rich features for better

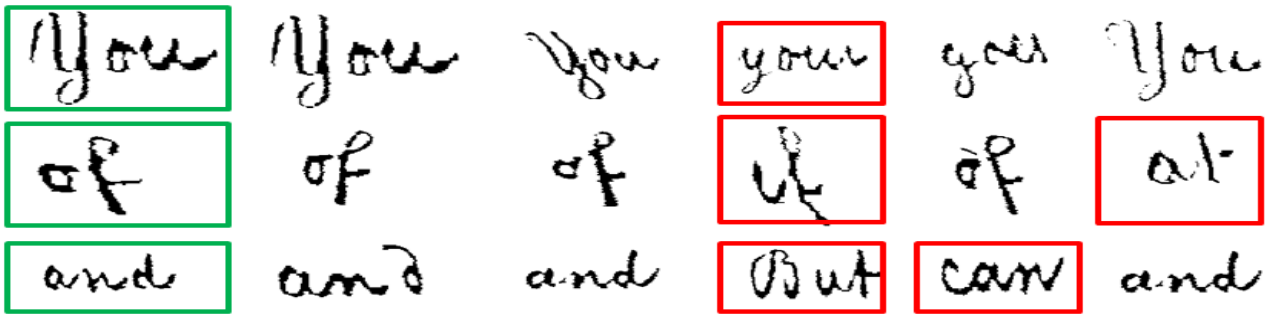

Fig. 2: Few sample results. Top-5 retrieval results for given query images. First column shows the query image. For each query, its top 5 retrieval images are shown from left to right. In each row, query image is marked in green colour and its corresponding wrong retrieval images are marked in red colour.

representation. However, in this experiment we use only simple profile features and does not apply any learning technique for obtaining better representation. The results are shown in Table III. We observe that our method achieves comparable performance with DTW based word image retrieval on the datasets of two Indian languages. This is intuitive because our approach is language independent, and hence is equally applicable to Indian languages.

We show the qualitative performance of proposed method on George Washington database in Figure 2. Here, we show some of the example queries and their top-5 retrieved word images. We have marked the query image in green colour and its incorrect retrieval imagesin red colour. It is to be specially noted that the current work does not use any learning feature models or any other post-processing techniques, and the main aim of this work is to show scalability of DTW based retrieval methods.

## VI. Conclusion

In this paper, we have proposed fast approximation of DTW distance. This is achieved by introducing global principal alignments, which avoids computation of optimal alignments for new query images. These precomputed global alignments captures all the correlations in the datasets. The proposed fast approximation of DTW distance makes DTW based document image retrieval computationally feasible. We have demonstrated utility of the proposed technique for retrieving word images from popular George Washington database and Indian language datasets. The performance of proposed method is as good as DTW distance and computationally performs equally as simple Euclidean based matching.

**Acknowledgements.** This work is supported by TCS PhD fellowship scheme.

## References

[1] John Aach and George M. Church. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 2001.

[2] Claus Bahlmann, Bernard Haasdonk, and Hans Burkhardt. On-line handwriting recognition with support vector machines - a kernel approach. In *Proc. of the 8th IWFHR*, 2002.

[3] A. Balasubramanian, Million Meshesha, and C. V. Jawahar. Retrieval from document image collections. In *DAS*, 2006.

[4] Michael K. Brown and Lawrence R. Rabiner. Dynamic time warping for isolated word recognition based on ordered graph searching techniques. In *ICASSP*, 1982.

[5] C. V. Jawahar and A. Kumar. Content-level annotation of large collection of printed document images. In *ICDAR*, 2007.

[6] G. Nagendar and C. V. Jawahar. Fast approximate dynamic warping kernels. In *ACM IKDD*, 2015.

[7] George Nagy. Twenty years of document image analysis in PAMI. *TPAMI*, 2000.

[8] Viresh Ranjan, Gaurav Harit, and C.V Jawahar. Document retrieval with unlimited vocabulary. In *IEEE Winter Conference on Applications of Computer Vision*, 2015.

[9] T. M. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 2007.

[10] Toni Rath and R. Manmatha. Lower-bounding of dynamic time warping distances for multivariate time series. In *MM*, 2003.

[11] Toni M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *CVPR*, 2003.

[12] David Sankoff and Joseph B. Kruskal, editors. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA, 1983.

[13] Hiroshi Shimodaira, Ken ichi Noma, Mitsuru Nakai, and Shigeki Sagayama. Dynamic time-alignment kernel in support vector machine. In *NIPS*, 2001.

[14] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[15] Xiaopeng Xi, Eamonn J. Keogh, Christian R. Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In *ICML*, 2006.

[16] Ismet Zeki Yalniz and R. Manmatha. An efficient framework for searching text in noisy document images. In *DAS*, 2012.

[17] Jian Yang, David Zhang, Alejandro F. Frangi, and Jing-Yu Yang. Two-dimensional pca: A new approach to appearance-based face representation and recognition. *PAMI*, 2004.