2009 10th International Conference on Document Analysis and Recognition

A rotation invariant page layout descriptor for document classification and retrieval

Albert Gordo, Ernest Valveny
Computer Vision Center
Universitat Autònoma de Barcelona
Spain
{agordo, ernest.valveny}@cvc.uab.es

Abstract

Document classification usually requieres of structural features such as the physical layout to obtain good accuracy rates on complex documents. This paper introduces a descriptor of the layout and a distance measure based on the cyclic Dynamic Time Warping which can be computed in $\mathcal{O}(n^2)$. This descriptor is translation invariant and can be easily modified to be scale and rotation invariant. Experiments with this descriptor and its rotation invariant modification are performed on the Girona Archives database and compared against another common layout distance, the Minimum Weight Edge Cover. The experiments show that these methods outperform the MWEC both in accuracy and speed, particularly on rotated documents.

1. Introduction

Document classification is an important task in document management and retrieval. It is usually based on the extraction of some features from the document image. These document features can be of different types. In [3] three categories are proposed (adapted from the four proposed in [2]): *image features*, extracted directly from the image or from a segmented image (e.g. the density of black pixels of a region), *structural features* or relationships between blocks in the page, obtained from the page layout, and *textual features*, based on the OCR output of the image. A classifier may combine these features to get better results.

Structural features are necessary to classify documents with structural variations within a class. For this task, several methods exist. Some of them define a distance between blocks and pages based on some criteria like the Minimum Weight Edge Cover [5] or the Earth's Mover Distance [9, 10]. Others represent the layout as a graph or tree and define some graph distances between them [2, 6]. Sometimes we can calculate the probability that a given graph is generated by a class representative graph generator [1]. Finally, documents can also be classified based on sets of rules [4].

In this paper we present a method to represent and classify document layouts based on a graph representation of the regions, which is later flattened into a cyclic sequence, obtaining a vector representation of the document layout. The comparison of these vector representations is done using the cyclic dynamic time warping. This approach allows a $\mathcal{O}(n^2)$ approximate comparison, which is fast compared to the typical exponential cost of some graphs methods or the $\mathcal{O}(n^3)$ of those based on the assignment problem. Moreover, this representation is invariant to translation and can be easily made invariant to scale and rotation.

In section 2 we introduce this representation along with the rotation invariant alternative. We also define the distance measures to compare two different page representations. Section 3 deals with the experimentation, and, finally, in section 4 we summarize the obtained results.

2. The cyclic polar page layout representation

The cyclic polar page layout is a cyclic sequence representation of the page layout. This approach uses an auxiliary complete bipartite graph, constructed with the centroids of the regions on one side and, on the other, the center of mass of all the regions. Figure 1 illustrates two sample pages with their layout and corresponding graph. Nodes of the graph are labeled with the area and the type of the region while edges are labeled with their length and either the angle or its delta increment. Then, the graph is converted into a vector representation obtaining a cyclic sequence. This sequence will later be compared by means of the cyclic dynamic time warping. This representation is translation invariant, and can be made scale invariant normalizing the mass and length. Moreover, it can be made rotation invari-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.110

ant disregarding the angles and keeping only their delta increments.

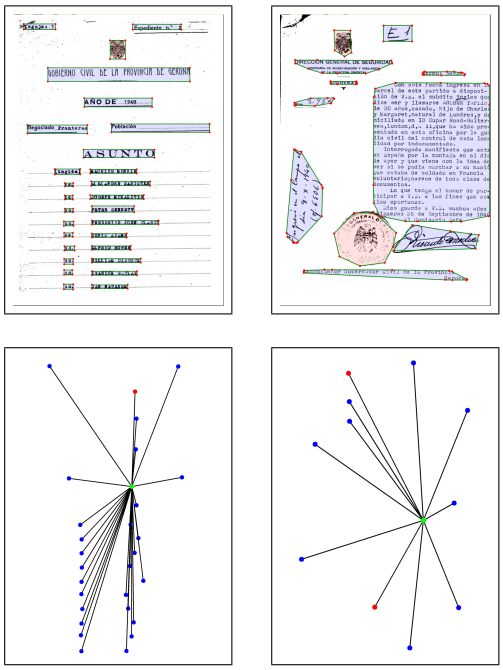

Figure 1. Two different layouts and their corresponding graphs.

2.1 Formulation

Let $Z = \{z_1, z_2, \ldots, z_m\}$ represent the m physical layout non-rectangular regions or zones of a segmented page, where each z is a 3-tuple $z_i = (C_i, A_i, T_i)$, C being the set of centroids of the zones, A their area and T their type (in our experiments, *text* or *non text*). Let also R be the center of mass of Z, where the mass M_i of z_i is equal to A_i.

Let the 3-tuple $G = (U, V, E)$ be a $K_{1,m}$ complete bipartite graph where $U = \{R\}$, $V = C$ and $E = U \times V$, and, particularly, $E = \{e_1, e_2, \ldots, e_m\}$ where $e_i = \overline{RC_i}$. Let the vertices be labeled with their mass (area) and type, and let the edges be labeled with their angle θ and their euclidean length L.

Let N be a sequential representation of G, $N = \{n_1, n_2, \ldots, n_m\}$ where n is a 4-tuple $n_i = (\theta_i, L_i, M_i, T_i)$ and where N is sorted as a function of θ.

A cyclic shift σ of a sequence $A = \{a_1, a_2, \ldots, a_m\}$ is a mapping $\sigma : \Sigma^* \rightarrow \Sigma^*$ defined as $\sigma(\{a_1, a_2, \ldots, a_m\}) = \{a_2, \ldots, a_m, a_1\}$. Let σ^k denote the composition of k cyclic shifts and let σ^0 denote the identity. Two sequences A and A' are cyclically equivalent if $A = \sigma^k(A')$, for some k. The equivalence class of A is $[A] = \{\sigma^k(A) : 0 \leq k < m\}$ and it is called a cyclic sequence.

Finally, let the equivalence class $[N]$ be our cyclic layout representation of the page.

2.2 Invariance

The above representation has the interesting feature of being translation invariant. The representation can also be made scale invariant normalizing the length and mass of N. Nonetheless, this should be avoided unless the databases actually contains significative scale variations within a class.

Also, replacing θ_i with Δ_i, where Δ_i is the absolute minimum angle formed by θ_i and $\theta_{(i+1) \bmod n}$ makes the representation rotation invariant. We will represent this invariant alternative as ΔN and its equivalence class as $[\Delta N]$. As we will prove through experimentation, this representation obtains better results on sets with rotations while keeping the good results on rotationless sets.

2.3 Distance measure

We must now define a distance measure between two layout representations $[N^a]$ and $[N^b]$ (or $[\Delta N^a]$ and $[\Delta N^b]$), cyclic sequences of sizes m and n. The very first step will be defining a distance D between sequences without considering the shiftings, i.e., between N^a and N^b. For this task, two obvious choices arise: the Edit Distance (ED) and the Dynamic Time Warping (DTW). The ED has the advantage of easily modelling some variations within a class like stamps or signatures that only appear sometimes and can be handled as insert or delete operations. On the other hand, DTW does a better job on common merge and split situations (due to errors in the layout segmentation or within class variations) thanks to the one-to-many correspondence. Preliminary experiments show that the DTW (coarsely normalized with the sum of the size of the sequences) offers better results than ED, and thus will be the one we will be using. For the DTW we must also define the cost function $\gamma(y, x)$ between nodes n_y^a and n_x^b. This can be simply defined as a linear combination of its parameters:

$$\begin{aligned}\gamma(y,x) = k_1 \text{ AngleDiff}(\theta_y^a, \theta_x^b)\ /\pi + \\ k_2 \text{ LengthDiff}(L_y^a, L_x^b) + \\ k_3 \text{ MassDiff}(M_y^a, M_x^b) + \\ k_4 \text{ TypeDiff}(T_y^a, T_x^b)\end{aligned} \quad (1)$$

where

- AngleDiff(θ_y^a, θ_x^b) is the absolute minimum difference of angles in radians.
- LengthDiff(L_y^a, L_x^b) is defined as $1 - \frac{L_y^a + L_x^b}{2 \max(L_y^a, L_x^b)}$

- MassDiff(M_y^a, M_x^b) is defined as $1 - \frac{M_y^a + M_x^b}{2\max(M_y^a, M_x^b)}$
- TypeDiff(T_y^a, T_x^b) is a binary function returning 0 if they belong to the same type and 1 if they do not.

If we are using the rotation invariant representation, AngleDiff(θ_y^a, θ_x^b) $/\pi$ should be replaced with DeltaDiff(Δ_y^a, Δ_x^b), which is defined as $1 - (\Delta_y^a + \Delta_x^b)/(2\max(\Delta_y^a, \Delta_x^b))$

Values for weights k_1 to k_4 can be obtained through validation as explained in section 3.

2.4 Cyclic distance measure

Once a distance between N^a and N^b has been defined we must define it for the equivalence classes $[N^a]$ and $[N^b]$. The exact distance can be naively computed with cost $\mathcal{O}(m^2n^2)$ as

$$CD([N^a],[N^b]) = \min_{\substack{0\le k<m \\ 0\le l<n}} D(\sigma^k(N^a), \sigma^l(N^b)) \qquad (2)$$

and in $\mathcal{O}(m^2 n)$ [8] as

$$CD([N^a],[N^b]) = \min_{0\le k<m} (\min D(\sigma^k(N^a), N^b), D(\sigma^k(N^a)N^a_{k+1}, N^b)) \qquad (3)$$

This last equation can also be solved in $\mathcal{O}(mn\log m)$ by means of a more elaborate divide & conquer approach [7, 8].

Also, suboptimal solutions can be found in $\mathcal{O}(mn)$, e.g., calculating the standard DTW between the concatenations of N^a and N^b with themselves, i.e.,

$$\text{ApproxCD}([N^a],[N^b]) = D(N^aN^a, N^bN^b) \qquad (4)$$

Measuring the distance with the cyclic DTW seems reasonable when using the rotation invariant descriptor $[\Delta N]$, but its usefulness when employing the angle-fixed $[N]$ is not that obvious. The reason why we use the cyclic DTW instead of the standard one is that on some descriptors there are many nodes with θ close to zero, above and below. Due to the cyclic nature of the angle difference, using the cyclic DTW we will allow some reasonable matches that were not possible with the standard DTW. This leads to better results even if the angles are fixed.

3 Experiments

To evaluate this cyclic representation of the layout we will test it with the Girona Archives database. The Girona database is a collection of documents from the Civil Government of Girona, in Spain, that contains documents related to people going through the Spanish-French border from 1940 up to 1976 such as safe-conducts, arrest reports, documents of prisoners transfers, medical reports, correspondence, etc. Even if it is a mostly-text database, in this case most of the pages also have images like stamps, signatures, etc in a non-manhattan disposition. We have used a subset of the database that contains 823 manually segmented images and is currently divided in 8 different categories. Some of these images are slightly skewed, but in most cases the skew is almost nonexistent. Some samples of the database can be seen in figure 2.

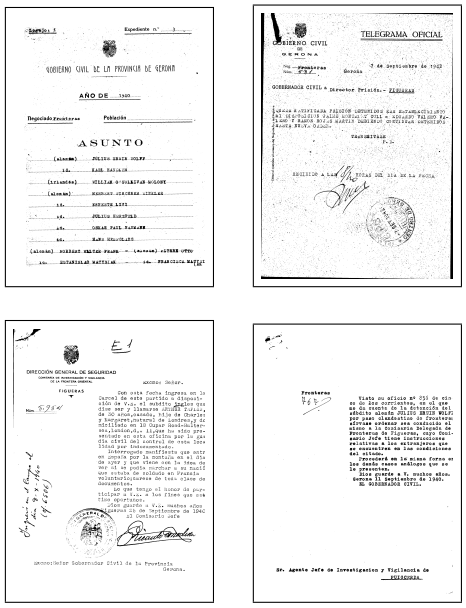

Figure 2. Samples of three different categories of the Girona database.

For all of our experiments the testing strategy will be leaving-one-out cross-validation. Experiments will be done both for classification, using a simple nearest neighbour classifier, and retrieval. First, we will try to obtain the best values for the weights k_1 to k_4, used in cost function of the approximate cyclic DTW (4) both with $[N]$ and $[\Delta N]$ representations. Later, we will compare the results to another common layout distance measure, the minimum weight edge cover. Finally, we will apply some random rotations to the pages and test how this affects the different distance measures.

3.1 Validation of cost function weights

As we explained in section 2.3, appropriate values for factors k_1 to k_4 in the γ cost function must be found. To do so, we will explore the space around the factors centered at $\{1, 1, 1, 1\}$. Table 1 shows the results for the three best classification error rates when using $[N]$, and table 2 does

the same for average precision rates in retrieval. Tables 3 and 4 show the same results for $[\Delta N]$.

k_1	k_2	k_3	k_4	Error rate
1.4	1.1	1.0	0.8	2.91
1.6	1.0	0.6	0.8	2.91
1.6	1.1	0.8	1.2	2.91

Table 1. Best error rates for $[N]$ as a function of factors k_1 to k_4

k_1	k_2	k_3	k_4	Average Precision
1.8	0.8	1.8	1.6	0.659
2.2	1.0	2.2	2.0	0.659
1.8	1.0	1.8	1.6	0.659

Table 2. Best average precision rates for $[N]$ as a function of factors k_1 to k_4

k_1	k_2	k_3	k_4	Error rate
0.2	1.2	1.0	1.0	3.76
0.8	1.3	1.0	1.0	3.76
0.9	1.2	1.0	1.0	3.76

Table 3. Best error rates for $[\Delta N]$ as a function of factors k_1 to k_4

k_1	k_2	k_3	k_4	Average Precision
1.3	1.2	1.2	1.0	0.642
1.2	1.2	1.2	1.0	0.642
1.1	1.2	1.2	1.0	0.642

Table 4. Best average precision rates for $[\Delta N]$ as a function of factors k_1 to k_4

In the case of $[N]$, we can see in table 1 that, for classification, all the weights are close to one. The most important weight is k_1, that is, the angle, with a slightly higher value. This leads to think that the absolute angle is important for classification, at least in $[N]$. However, in the case of $[\Delta N]$, the results are quite different; the angle difference seems to be less significant, as its weight has a high variance while the rest of the weights remain almost constant (table 3).

When dealing with the long term retrieval problem, one would intuitively expect all factors to have approximately the same weight. This is true for $[\Delta N]$ and almost true for $[N]$, where the edge length has less importance than the rest of weights (tables 4 and 2 respectively).

3.2 Comparison with another layout distance

To test the results of this representation, we will compare our results against another common layout distance, the Minimum Weight Edge Cover (MWEC) [5]. This distance is based on the assignment problem and has an asymptotic cost of $\mathcal{O}(n^3)$. It has also been proven to provide better results than other similar measures like the pure assignment problem or the earth's mover distance [10].

Table 5 shows the best results obtained in classification and retrieval with the approximate cyclic DTW over $[N]$ and $[\Delta N]$ compared to those obtained using the MWEC along with their average time[1]. We can see that $[N]$ obtains better results than MWEC both in classification and retrieval. In the case of $[\Delta N]$, the results do not seem that good as its classification rate is the worst of them all. Nonetheless, the long term retrieval is almost five points above the results obtained with MWEC spending a third of the time.

	Error Rate	Average Precision	Av. Time
$[N]$	2.91	0.659	5.0s
$[\Delta N]$	3.76	0.642	5.2 s
MWEC	3.64	0.598	16.0s

Table 5. Classification and retrieval rates with different tuned distances.

It should be noted that a deep inspection of the results reveals that most of the $[\Delta N]$ and $[N]$ errors are produced within two categories whose pages contain a very low number of zones. In this case, the translation and rotation invariance is inconvenient, as this is a critical information for the correct classification. Moreover, in pages with few zones, the center of mass is subject to a high variance, leading to inappropriate descriptions. On the other hand, the MWEC correctly classifies these regions most of the time. A combined classifier using MWEC when the page contains few zones and $[N]$ or $[\Delta N]$ otherwise would most likely outperform any of the methods separately.

3.3 Rotation

In this last experiment we will observe how the cyclic DTW over $[N]$ and $[\Delta N]$ compares to the MWEC when the

[1]Time to compare each of the 823 layouts against each other

pages can be rotated. To do so, we will first apply to each page an uniform random rotation in the range $[-\pi/2, \pi/2]$ radians and we will re-learn the appropriate weights of the γ function for $[N]$ both for classification and retrieval. As $[\Delta N]$ is supposed to be rotation invariant, in this case we will use the weights previously obtained in section 3.1. Then we will apply a second, different rotation to all the pages and we will check the results using $[N]$ with the new weights, $[\Delta N]$ and MWEC.

The best learned values for $[N]$ can be seen in table 6 ($[N]^c$ for classification and $[N]^p$ for retrieval). We can see that, as expected due to the large variability in angle introduced by the rotation, the angle weight is *much* less important than the rest of factors. It seems like the angle has a low importance, but it should not be forgotten that the whole structure would keep information about the radial order of the zones even if this weight was set to zero. It is also interesting that, even in this case, the retrieval precision is higher than the one obtained with the MWEC.

	k_1	k_2	k_3	k_4	E.R.	A.P.
$[N]^c$	0.1	0.9	1.6	1.0	4.25	0.600
$[N]^p$	0.1	1.3	1.4	1.4	5.10	0.606

Table 6. Best $[N]$ values for factors k_1 to k_4 with rotated pages.

Table 7 shows the results of classification and retrieval over the second set of rotated pages. As expected, the results obtained with $[\Delta N]$ are better than those obtained either with $[N]$ or MWEC. Nonetheless, the results obtained with $[N]$ are quite close to those obtained with $[\Delta N]$ as the angle weight is set to almost zero. It should also be noted that the results of $[\Delta N]$ are not exactly equal to those obtained previously. This difference is due to the fact that we are using an approximate cyclic DTW and not an exact one.

	Error Rate	Average Precision
$[N]$	5.34	0.612
$[\Delta N]$	4.98	0.624
MWEC	10.20	0.380

Table 7. Classification and retrieval rates over a set of randomly rotated pages.

4 Conclusions

In this paper we have presented a descriptor for page layouts and a distance measure between page descriptions that can be computed in $\mathcal{O}(n^2)$. This descriptor is translation invariant and can be modified to be scale and rotation invariant. Experiments with the non-manhattan Girona Archives database prove these methods to perform better than the common Minimum Weight Edge Cover both in the classification and retrieval and in speed. Most of the classification errors are produced on documents with few zones, so the use of this descriptor is not encouraged in such cases and an hybrid method is suggested.

When the documents are rotated, the rotation invariant descriptor clearly outperforms the MWEC. The rotation invariant descriptor performs only slightly better than the unmodified one when its angle weight is set to a low value, which leads to think that the exact angle is not as important as the overall structure. Finally, using the exact cyclic DTW would yield better results both in classification and retrieval using $[\Delta N]$ over rotated images, but the computation cost would increase.

References

[1] A. D. Bagdanov and M. Worring. First order gaussian graphs for efficient structure classification. *Pattern Recognition.*, 36(6):1311–1324, 2003.

[2] F. Cesarini, M. Lastri, S. Marinai, and G. Soda. Encoding of modified x-y trees for document classification. *Proceedings. Sixth International Conference on Document Analysis and Recognition*, pages 1131–1136, 2001.

[3] N. Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *Int. J. Doc. Anal. Recognit.*, pages 1–16, 2007.

[4] F. Esposito, D. Malerba, and F. A. List. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems.*, 14(2):175–198, 2000.

[5] D. Keysers, T. Deselaers, and H. Ney. Pixel-to-pixel matching for image recognition using hungarian graph matching. *In DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, pages 154–162, 2004.

[6] J. Liang, D. Doermann, M. Ma, and J. Guo. Page classification through logical labelling. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 3:477–480 vol.3, 2002.

[7] M. Maes. On a cyclic string-to-string correction problem. *Inf. Process. Lett.*, 35(2):73–78, 1990.

[8] A. Marzal and V. Palazón. Dynamic time warping of cyclic strings for shape matching. *Pattern Recognition and Image Analysis*, pages 644–652, 2005.

[9] Y. Rubner, L. Guibas, and C. Tomassi. The earth mover's distance, multi-dimensional scaling , and color-based image retrieval. *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668, May 1997.

[10] J. van Beusekom, D. Keysers, F. Shafait, and T. Breuel. Distance measures for layout-based document image retrieval. *Document Image Analysis for Libraries, 2006. DIAL '06. Second International Conference on*, pages 11 pp.–242, April 2006.