필기 문자의 확률 분포 유사성에 기반한 필기자 종속적 온라인 필기 한글 낱자 생성

최현일O 김자환 김진형

한국과학기술원 전자전산학과
{hichoiO, jahwan, jkim}@ai.kaist.ac.kr

Writer Dependent Online Hangul Syllable Generation based on Similarity between Probabilistic Distributions of Handwritings

Hyunil Choi0 Jahwan Kim Jin Hyung Kim

CS Div., EECS Dept, KAIST

요 약

필기는 특정인의 정체성을 나타내는 매우 효과적인 매체이다. 이런 이유로 최근 개인의 필기를 디지털 환경에서 사용하기 위한 방법들이 제안되고 있다. 또한 필기 인식 분야에서는 높은 성능의 필기 인식 시스템을 개발하기 위해 많은 양의 필기 데이터를 필요로 한다. 이에 본 연구는 주어진 데이터내의 문자 조합을 바탕으로 한글 필기를 생성하는 것을 목적으로 한다. 특히 데이터에서 관측되지 않은 필기까지도 생성하는 것을 주요 문제로 다룬다. 실험 결과 생성된 필기는 실제 필기와 시각적으로 매우 유사함을 보인다.

1. 서 론

개인의 필기는 자신의 정체성을 나타내는 효과적인 매체이다. 이는 필기자의 신경 근육학적 요인뿐만 아니라 심리학적, 환경적 요인도 필기 생성 과정에 융합되기 때문이다. 최근 개인 필기가 나타내는 정체성을 바탕으로 [1] 과 같이 디지털 환경에서 필기자의 개성을 나타내는 응용 분야가 창출되고 있다. 또한 이런 개인 필기 성향은 높은 성능의 필기 인식 시스템을 개발하는데 필수적으로 고려되어야 할 사항이다. 그러나 현실적으로 필기자로부터 많은 양의 데이터를 구하기가 어렵다. 이에 본 연구는 주어진 데이터내의 문자를 조합함으로써 한글 필기를 생성하는 것을 목적으로 한다. 특히 데이터에서 관측되지 않은 필기까지도 생성하는 것을 주요 문제로 다룬다. 개발된 시스템은 개인 필기의 글꼴화 및 가상 필기 데이터의 생성 등에 사용될 수 있을 것이다.

필기 생성에 관한 기존 연구로는 필기 생성 과정의 운동학적 모델링[2, 3] 방식, 주어진 데이터의 변형을 통한 생성 [4, 5, 6]등이 있다. 운동학적 모델링 중 대표적인 방법으로서 delta-lognormal theory[2] 는 펜 끝의 속도를 제어하는 함수들을 운동학에 기반하여 정의하고 이의 파라미터 값을 조정함으로써 다양한 필기를 생성하였다. 주어진 데이터의 변형을 통한 생성 방식은 문자에 대한 구조를 정의하고 이의 파라미터 값을 데이터로부터 추정한 후, 미리 정의된 과정에 의해 필기를 생성하는 것이다.

대부분의 관련 연구에서 다루고 있는 문제는 “주어진 데이터 혹은 template 의 변형” 에 초점을 맞추고 있다. 이 방법론은 어떤 패턴의 변형을 생성하기 위해서는 반드시 그를 위한 데이터가 존재해야 한다는 관점에서 타당하다. 그러나 영어 단어에 해당하는 한글 낱자나 한문 등에는 매우 많은 수의 클래스가 존재하기 때문에 특정 필기자로부터 모든 클래스에 대한 데이터를 모으기가 현실적으로 매우 어렵다.

본 논문에서는 한글의 구조적 특성을 바탕으로 데이터에서 관측된 문자뿐만 아니라 관측되지 않은 문자 까지도 생성 하는 방법을 제안한다. 관측된 문자의 경우, 기존 방법들[4, 5]을 통해 복수개의 문자들로부터 변이를 생성할 수 있다. 관측되지 않은 문자는 문자에 대한 어떠한 정보도 없기 때문에 생성하기가 매우 어렵다. 우리는 이를 위해 개인 필기도 필기자 공통적인 필기 경향을 따른 다는 것을 전제로 필기자 공통적인 필기 형태를 확률적으로 모델링 한 후, 이들간의 유사성을 바탕으로 관측되지 않은 문자를 관측된 문자들로 조합한다. 문자 조합의 최소 단위로서 자모를 이용하며 낱자 내에서 이의 형태 및 크기, 위치를 모델링 한다. 데이터에서 관측되지 않은 자모 형태 및 크기 위치는 모델에서 확률이 최대로 되는 인스탄스를 생성함으로써 구해진다.

논문의 순서는 다음과 같다. 2 절에서는 제안하는 시스템의 개요를, 3 절에서는 문자의 확률적 모델링을 보인다. 문자 모델과 주어진 데이터로부터 어떻게 필기를 생성하는 가는 4 절에서 다루어진다. 5 절에서는 실험 결과를, 마지막으로 6 절에서는 결론 및 향후 연구를 서술한다.

2. 시스템 개요

시스템의 입력은 특정 필기자의 한글 낱자와 해당하는 레이블(label)로 구성된 데이터 D, 생성하고자 하는 한글 레이블 l_t 로 구성된다. 모든 레이블은 자모 레이블로 분리될 수 있으며 각 자모들에 해당하는 점들의 집합 $(x_1 \mathrm{K}\ x_M)$ 을 가진다[1]. 기본적으로 l_t 에 대한 생성은 데이터에서 적절한 자모를 선택함으로써 이루어진다. l_t 의 생성에 필요한 자모가 데이터에서 관측되었는지의 여부에 따른 간략한 필기 생성 규칙이 그림 1에 나타나 있다.

if $l_t = l_i \in D$ and all graphemes seen **then**
  generate a variant of $\boldsymbol{x}_i$ or reproduce $\boldsymbol{x}_i$
else if $l_t = l_i \notin D$ and all graphemes seen **then**
  synthesize $\boldsymbol{x}_t$ with corresponding $\boldsymbol{x}_i$
else if $l_t = l_i \notin D$ and some graphemes seen, some graphemes unseen **then**
  synthesize $\boldsymbol{x}_t$ with corresponding $\boldsymbol{x}_i$ and predict unseen graphemes
else if $l_t = l_i \notin D$ and all graphemes unseen **then**
  generate most probable $\boldsymbol{x}_t$ from writer independent data
end if

그림 1. 자모의 관측 여부에 따른 필기 생성 규칙

이와 같은 규칙을 통해 제한된 한글 낱자 집합으로부터 모든 가능한 한글 필기를 생성할 수 있게 된다. 그림 1에서 첫 번째 조건은 기존의 필기 생성 연구들이 주로 접근했던 문제이고 나머지 조건들은 본 연구에서 중점적으로 풀고자 하는 문제들이다. 우리는 이 규칙하에서 다음과 같은 문제들을 생각해볼 수 있다 1) l_t 내 자모가 관측 되었다면 어떤 기준으로 해당되는 낱자의 자모를 선택할 것인가 2) l_t 내 자모가 관측 되지 않았다면 어떻게 자모의 형태를 예측할 수 있는가?

두 가지 문제를 해결하기 위해 우리는 낱자 내 자모들 사이의 형태적/위치적 관계를 나타내는 *필기자 공통적인 확률 분포*가 있다고 보고 이의 유사성을 바탕으로 l_t 의 자모들을 데이터에서 선택한다. l_t 의 자모들이 관측되지 않았을 경우, 해당되는 낱자의 확률 분포 상에서 확률을 최대로 하는 인스탄스를 생성함으로써 l_t 의 자모들을 예측 할 수 있다. 이러한 확률 분포는 필기자 독립 데이터로부터 추정된다.

3. 문자 형태의 확률적 모델링

문자의 형태를 확률적으로 모델링 함에 있어서 확률 변수를 어떻게 정의할 것인가는 매우 중요한 문제이다. 최근에 [7, 8] 등의 연구에서 나타났던 획 간의 관계 모델링은 높은 성능의 인식을 수행하는 결과를 보였다. 이는 획이 가지고 있는 문자 형태의 표현력과 이들 사이의 확률적 관계가 필기의 변이를 매우 효과적으로 표현한다는 것이라고 말 할 수 있다. 본 연구도 획을 기반으로 문자의 형태를 표현한다.

어떤 한글 낱자 H 는 여러 개의 자모 G 로 구성 되며 각각의 자모는 여러 개의 획으로 구성된다. 한글 낱자의 결합 확률 분포는 필기 순서에 의해

$$p(H_l) = p(G_1^l \mathrm{K}\ G_M^l) = \prod_{i=1}^{M} p(G_i^l \mid pa(G_i^l)) \qquad (1)$$

로 정의된다.

여기서 $pa(G_i^l)$ 는 G_i^l 이전 자모들의 집합이며 이는 G_i^l 에 영향을 미치는 자모들이다. 여기서 N 개의 조건부 확률 분포는 조건부 가우시안 확률 분포[9]로 표현된다.

우리는 자모의 크기 및 위치의 유사성 또한 생성시 중요한 속성이 된다는 것을 관찰하였다. 이를 위해 자모의 형태는 획으로, 자모의 문자 내 크기 및 위치는 자모의 외곽사각형(bounding box) 으로 분리하여 모델링 한다.

4. 필기 생성 알고리즘

우리는 주어진 데이터에 대응되는 문자의 확률 분포를 3절에서 구했다. 이는 생성하고자 하는 문자를 데이터에서 어떻게 선택하는가 또는 예측하는가에 대한 기준을 제공한다. 이 때 필기자 독립적인 확률 분포만을 이용하게 되면 개인 필기에 존재하는 변이를 반영하지 못하게 된다. 이를 위해 데이터에서 관측된 자모가 주어졌을 때의 조건부 확률 분포를 이용한다.

어떤 두 문자 l_1, l_2 의 확률 분포를 f^{l_1}, f^{l_2} 라고 할 때, i 번째 자모 $G_i^{l_1}$, $G_i^{l_2}$ 의 분포의 유사성을 측정하는 방법은 관측된 자모가 주어졌을 때 각 문자의 조건부 확률 분포의 유사성을 바탕으로 한다.

$$\begin{aligned}\mathrm{dist}(&f^{l_1}(G_i^{l_1} \mid G_{i-1}^{l_1} = g_{i-1}^{l_1} \mathrm{K}\ G_1^{l_1} = g_1^{l_1}),\\ &f^{l_2}(G_i^{l_2} \mid G_{i-1}^{l_2} = g_{i-1}^{l_2} \mathrm{K}\ G_1^{l_2} = g_1^{l_2}))\end{aligned} \qquad (2)$$

위에서 dist 는 두 분포의 유사 정도를 측정하는 거리 함수이며 값이 0 에 가까워질수록 두 분포는 유사하다[2]. 개인 필기 데이터 g_i^l 은 G_i^l 에 대한 인스탄스이다.

관측되지 않은 문자 l_t 내의 어떤 자모 $g_i^{l_t}$ 를 $C = \{g_i^{l_1} \mathrm{K}\ g_i^{l_K}\}, l_t \notin \{1\mathrm{K}\ K\}$ 에서 선택해야 할 때 식 (2)

[1] 필기 인식을 통해 자모에 해당하는 점들의 집합을 구할 수 있다.

[2] dist 는 상대적 엔트로피(relative entropy) 로 정의되었다.

는 f^{l_t} 와 $f^{l_k}, l_k \in \{l_1 \mathrm{K}\ l_K\}$ 의 dist 값이 최소가 되는 문자 내 자모를 선택한다. 이 때 $g_{i-1}^{l_t} \mathrm{K}\ g_1^{l_t}$ 은 식 (2) 가 재귀적으로 적용되었을 때 선택된 값들이다. 만약 $C = \phi$ 일 때는 다음 식에서 확률이 최대가 되는 $g_i^{l_t}$ (MPE, Most Probable Explanation) 를 선택한다.

$$g_i^{l_t} * = \arg\max_g p^{l_t}(G_i^{l_t} = g \mid G_{i-1}^{l_t} = g_{i-1}^{l_t} \mathrm{K}\ G_1^{l_t} = g_1^{l_t}) \tag{3}$$

식 (3) 에서 곡선 형태의 획 대한 MPE 를 구하기 위해 [7] 에서 제안된 재귀적 획 분할 기법을 이용한다.

5. 실험결과

문자의 확률 분포는 KAIST 한글 데이터베이스로 추정되었다. 필기 생성을 위해 8 명의 필기자로부터 5개의 문단으로 이루어진 307 개의 낱자를 5번씩 입력 받았으며 모두 144 개의 클래스로 구성된다[3].

제안된 방법의 평가는 문단 단위 cross validation 을 바탕으로 시스템이 생성한 필기와 실제 필기 사이의 시각적 분석을 수행함으로써 이루어진다. 다음 그림은 텍스트의 문단 1-4 로부터 문단 5 를 생성한 결과 중 일부분을 보이고 있다. 첫 번째 줄은 실제 필기 데이터를, 두 번째 줄은 시스템이 생성한 필기를 나타낸다. 각 열은 해당 문자가 어떤 글자의 자모들로부터 생성되었는지를 나타낸다.

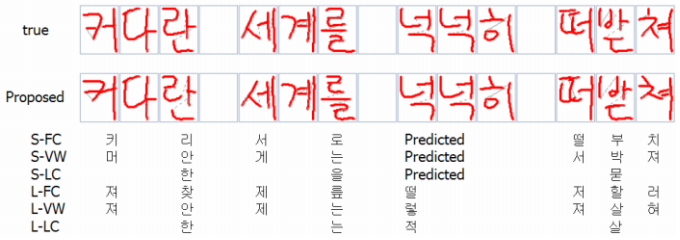
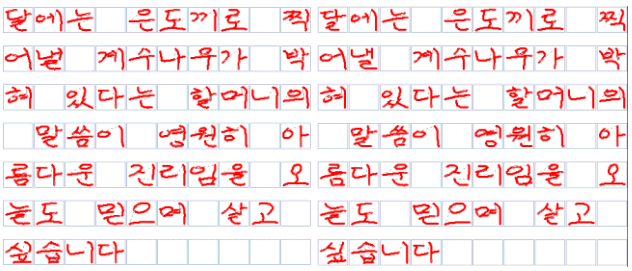


그림 2. S 는 자모 형태를, L 은 자모의 크기 및 위치를 나타내고 FC, VW, LC 는 각각 초성, 중성, 종성을 의미한다. Predicted 는 확률 분포로부터 예측된 자모 이다.

위의 결과에서 대부분의 생성 결과가 시각적으로 실제 데이터와 유사성을 보이고 있음을 알 수 있다. 또한 선택된 자모들은 생성하고자 하는 문자와 구조적인 유사성을 보이고 있다. 예를 들어, "란" 의 경우 선택된 후보들은 1) 초성: 중성의 시작 위치, 2) 중성: 동일한 종성, 3) 종성: 동일한 중성 의 특성을 보인다. 다른 예로, "받" 의 경우, 초성과 종성은 이웃하는 자모의 배치 구조가 상이한 문자에서 선택되었다. 그럼에도 불구하고 자연스러운 문자 형태를 보이고 있다. 이는 데이터에 구조적 유사성을 가지는 후보가 존재하지 않더라도 확률적으로 가장 유사한 문자를 선택하려는 시스템의 특징을 잘 보이고 있다.



(a) (b)

그림 3. 필기 생성 예 (a) 실제 필기 (b) 합성된 필기

6. 결론 및 향후 연구

본 연구는 문자 확률 분포의 유사성을 바탕으로 데이터 내의 자모 조합을 통해 한글 필기를 생성하였다. 특히 데이터에서 관측되지 않은 필기를 생성했을 때도 생성된 필기는 실제 필기와 시각적 유사성을 보였다. 향후 연구로는 문자 확률 분포의 필기자 적응을 통한 생성 등이 있다.

참고 문헌

[1] 스타폰트, http://www.hanyang.co.kr, 한양시스템

[2] R. Plamondon and W. Guerfali, The generation of handwriting with delta-lognormal synergies, Biological Cybernetics, 78, pp. 119-132, 1998

[3] D.-H. Lee and H.-G. Cho, A New Synthesizing Method for Handwriting Korean Scripts, The Visual Computer Journal, 17(3):147-157, 2001

[4] J. Wang, Chenyu Wu, Ying-Qing Xu, Heung-Yeung Shum, and Liang Ji, Learning-based cursive handwriting synthesis, International Workshop on Frontiers in Handwriting Recognition, pp. 157-162, 2002

[5] H. I. Choi, S. J. Cho, and J. H. Kim, Writer dependent online handwriting generation with bayesian network, International Workshop of Frontiers in handwriting Recognition, pp. 130-135, 2004

[6] A. Herztmann, N. Oliver, S. Seitz, and B. Curless, Curve analogies, Eurographics Workshop on Rendering, pp. 233-246, 2002

[7] S. J. Cho and J. H. Kim, Bayesian network modeling of strokes and their relationships for on-line handwriting recognition, Pattern Recognition, 37(2), pp. 253-264, 2004

[8] I. J. Kim and J. H. Kim, Statistical Character Structure Modeling and Its Application to Handwritten Chinese Character Recognition, IEEE TPAMI, 25(11), pp. 1422-1436, 2003

[9] K.P. Murphy, Inference and learning in hybrid Bayesian networks. Technical Report 990, U.C.Berkeley, Dept. Comp. Sci, 1998

[3] 본 논문에 사용된 텍스트는 정한모의 "가을에" 이다.