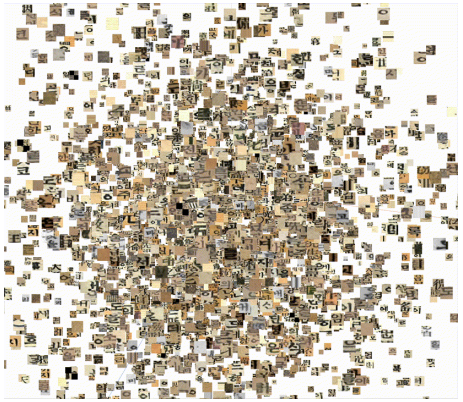
Visual Feature Discovery in Colonial Korean Print using MIL

DHBenelux 2025



Aron van de Pol, Jelena Prokic, Angus Mol

Leiden University Cente for Digital Humanities & Department of Korean Studies

2025-06-14

March 1st 1919

Image source: Schofield[1]

1. "Image of Crowd Outside of City Hall Taken by Dr. Schofield on March 1, 1919".

February 27th 1919

- In the evening Ch'oe Nam-sŏn 崔南善 typset the Document at the *Sinmungwan* 新文舘
- He then transferred it to Yi Chong-il 李鍾一 at the *Posŏngsa* 普成社
- That night, the Posŏngsa produced 21.000 copies[1]

Figure 1: The Sinmungwan.

Figure 2: The Posŏngsa.[2]

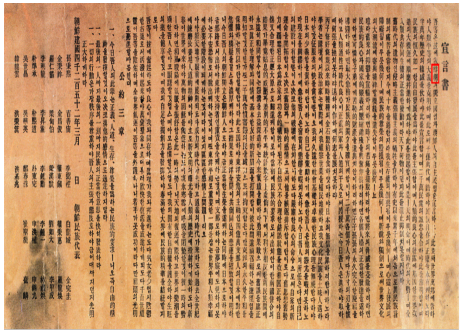


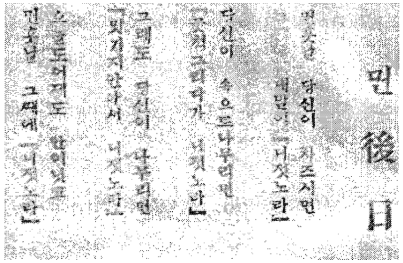
Figure 3: The Declaration of Independence. Minjok Taep'yo 33-in 民族代表 33人[3]

Figure 4: Example of a typeset document *Chindallaekkot* 진달래꽃.[4]

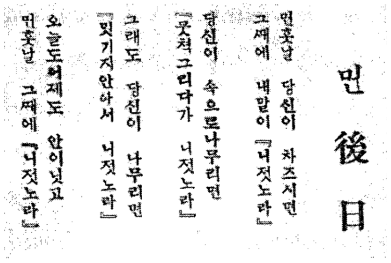
1. For a more detailed overview see: Shin, *Korean National Identity Under Japanese Colonial Rule*; Park, "3.1 Tongnip Sŏnŏnsŏ Inswae Kwajŏng kwa P'anbon üi Kŏmt'o 3.1독립선언서 인쇄과정과 판본의 검토 [A Review of the Printing Process and Editions of the Korean Declaration of Independence in 1919]".
2. The building can be seen on the top right. The building in the center housed the Sinchukgyosa

The 1920s

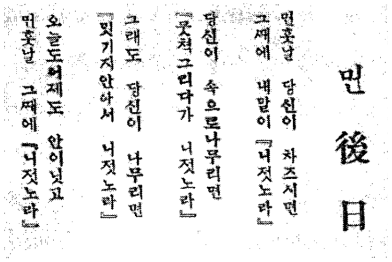
- During the 1920s, choice of 100-200 different print shops.[1]
- Choice influenced outcome.
- A good example are poems from Kim So-wŏl's *Chindallaekkot* 진달래꽃 collected works.



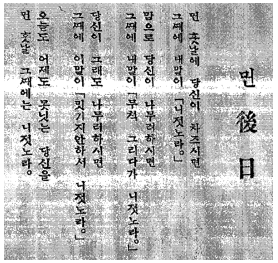
(a) Hansong Toso 漢城圖書 *issue of Chindallaekkot (collected works)*



(b) Chungang Sorim 中央書林 *issue of Chindallaekkot (collected works)*



(c) Haksaenggye 學生界 *(July 1920) issue of the poem.*

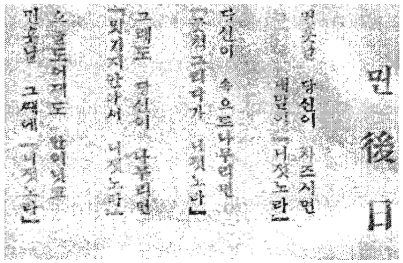


(d) Kaebyok 開闢 *(August 1922) issue of the poem.*

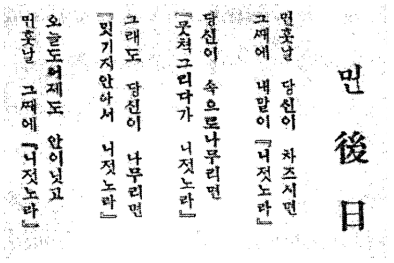
Figure 5: *Some Day Long From Now* 먼 후일 poem from various print runs

The 1920s

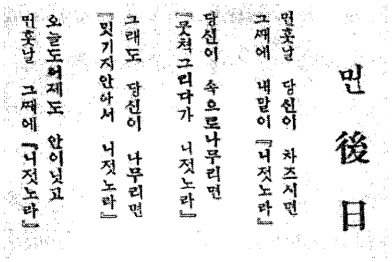
- During the 1920s, choice of 100-200 different print shops[1]
- Choice influenced outcome
- A good example are poems from Kim So-wŏl's *Chindallaekkot* 진달래꽃 collected works



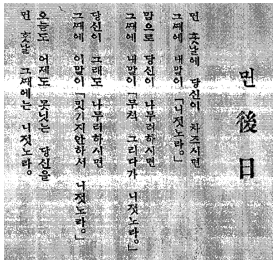
(a) Hansong Toso 漢城圖書 issue of Chindallaekkot (collected works)



(b) Chungang Sorim 中央書林 issue of Chindallaekkot (collected works)*



(c) Haksaenggye 學生界 (July 1920) issue of the poem.



(d) Kaebyok 開闢 (August 1922) issue of the poem.

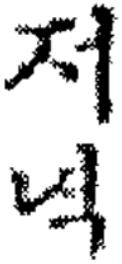
Figure 6: *Some Day Long From Now* 먼 훗날 poem from various print runs.

Characteristics?

- Work done by De Fremery[1]

(a) Hansong Toso Chusik Hoeisa printed Tang 당

(b) Taedong Inswaeso printed Tang 당



(c) Hansong Toso Chusik Hoeisa printed Tang 당

(d) Taedong Inswaeso printed Tang 당

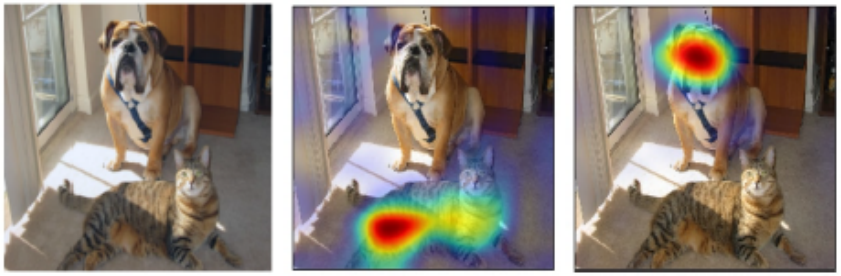
Figure 7: Examples of subtle typographic differences. Notice the angle of the stroke on *Tigŭt* ㄷ and the differences in the *Chiŭt* ㅈ

Research

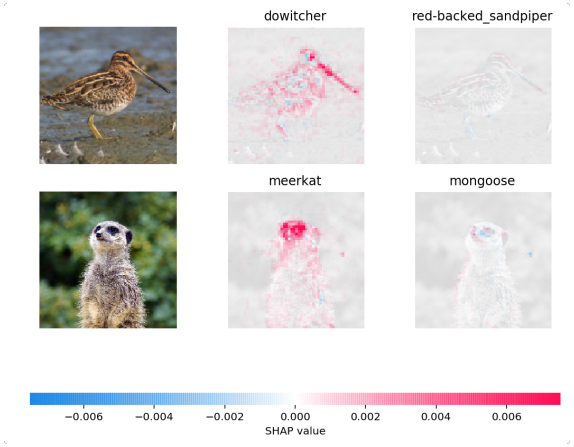
> Can neural networks be used to classify historical printshops and identify the specific visual features that distinguish their typographic styles?

Interpretability

- All studies aim to detect, but *how* a model detects is often neglected
  - For CNNs and Vision Transformers, several interpretability methods have proven successful
  1. GradCAM and derivates[1]
  2. SHAP (SHapley Additive exPlanations)[2]



(a) Example of GradCAM



(b) Example of SHAP


Figure 8: Interpretability methods visualized

Dataset

- We scraped the Hyundam Mun'go for magazines & paperbacks dating between 1900-1950.[1]
- 177.101 Images of Pages.
- 14.597 Publications.
- Contributions of 202 unique print shops.
- 2552 Publishers.
- 787 Distribution outlets.

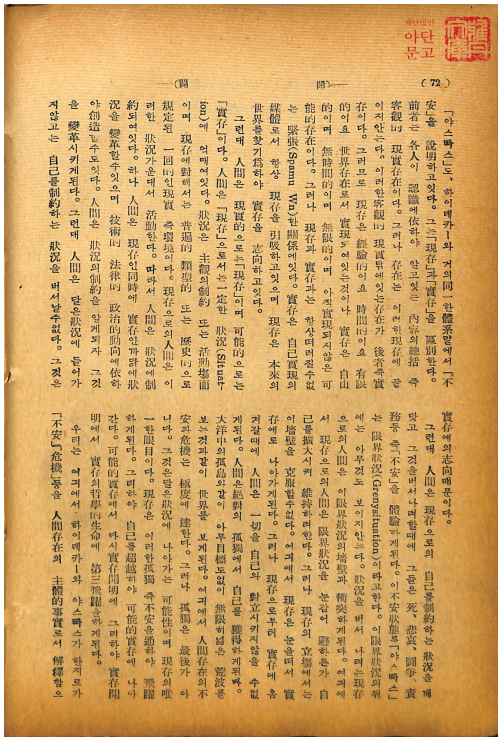
Note: These are uniquely named entities. The Publishers/ Distribution Outlets need more cleaning for a definitive answer.

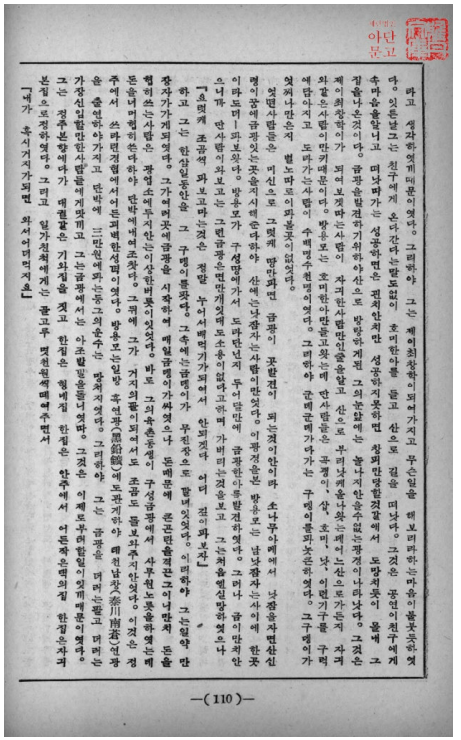
Dataset

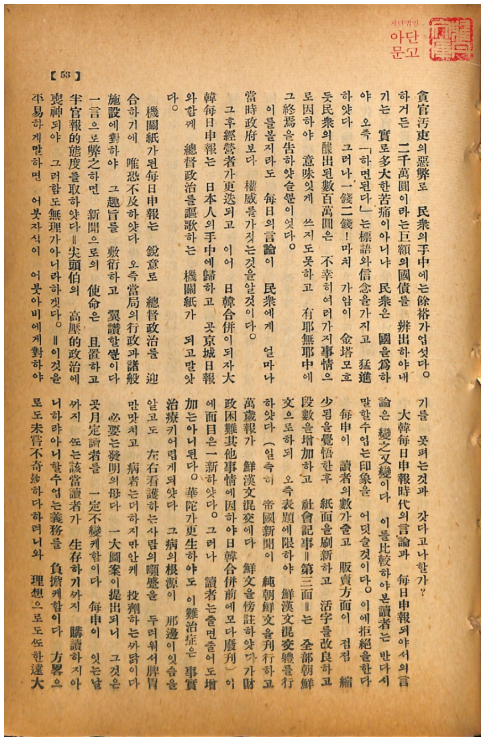
Printshop	Printshop (KR)	Pages	Percentage
Taedong Inswaeso	大東印刷所	27,882	42.07%
Hansŏng Toso Chusik Hoeisa	漢城圖書 株式會社	19,244	29.04%
Sinmungwan	新文館	13,050	19.69%
Chosŏn Inswae Chusik Hoeisa	朝鮮印刷株式會社	6,101	9.20%
Total		**66,277**	**100%**

- **Class Imbalance:** This distribution reflects real-world production volumes.
- We Chose not to implement class rebalancing techniques.

Dataset



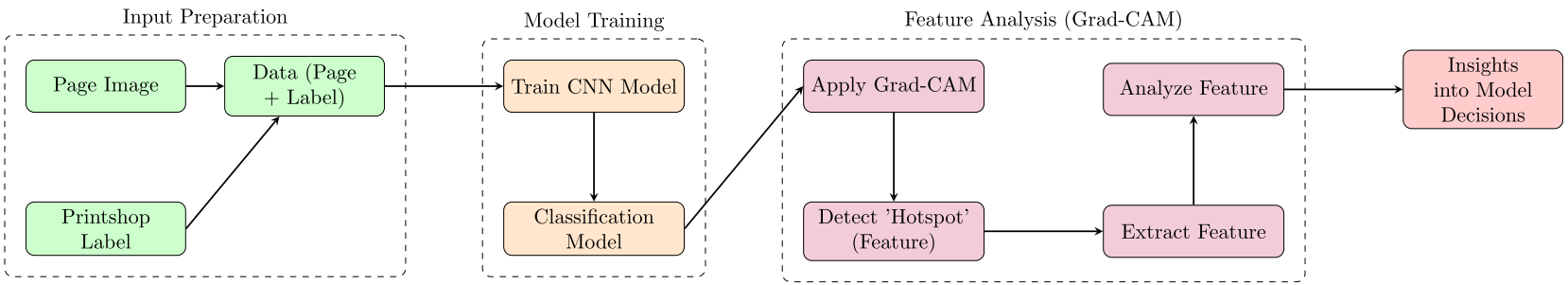




Examples of pages in the dataset.

Approach 1

CNN/ViT[1] + GradCAM

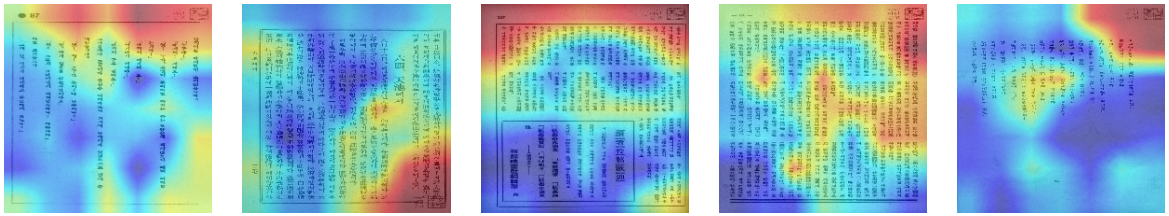


Results Approach 1

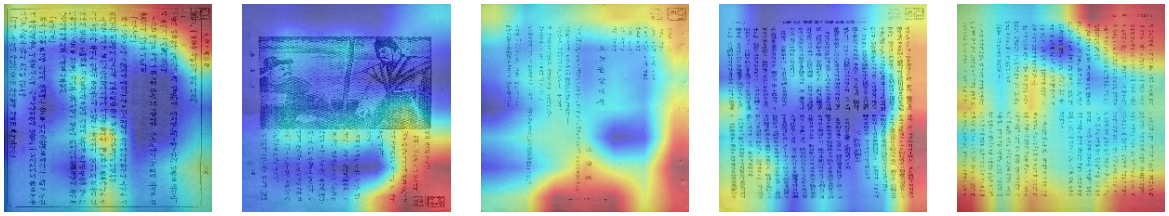
ConvNext Base architecture - 98% Accuracy (F1=0.98)

Chosŏn Inswae Chusik Hoeisa

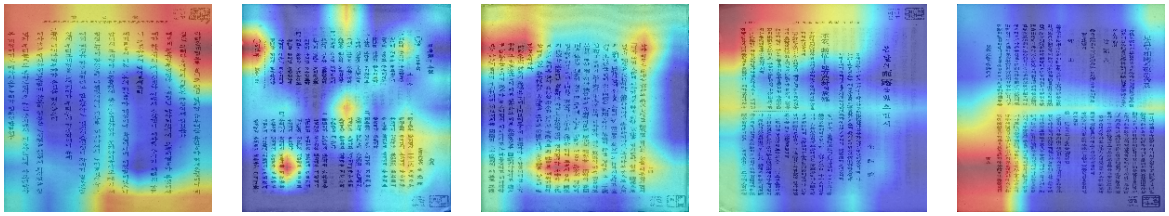
Hansŏng Toso Chusik Hoeisa



Taedong Inswaeso

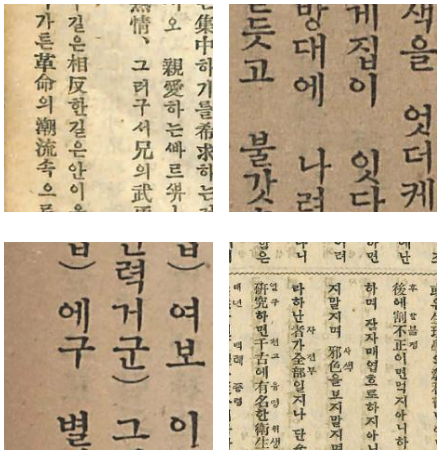


Sinmungwan



Approach 2

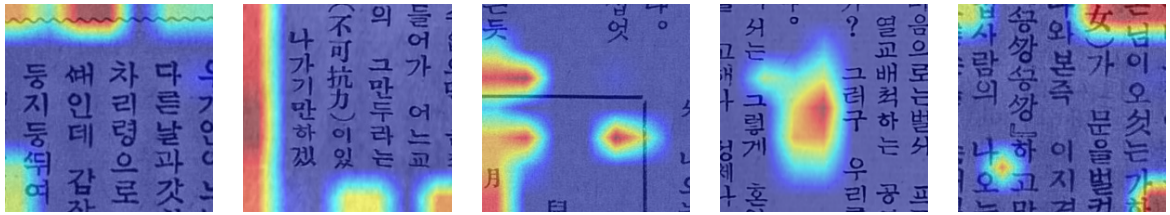
Following idea of Seuret et al.[1] a page is cut into 4 random cutouts, while reducing overlap to max 30%



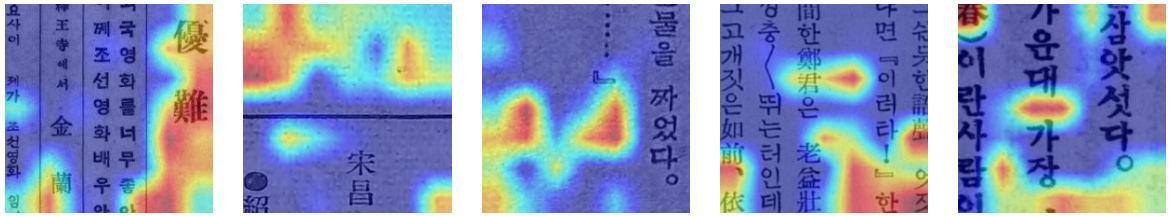
Approach 2 Results

99.8% Accuracy (F1=0.99) Swin S3 Base-224

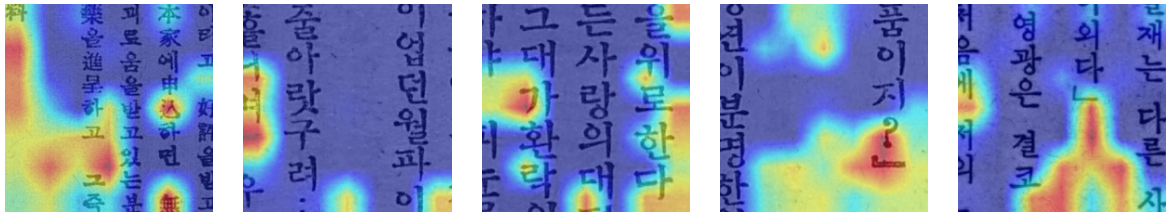
Chosŏn Inswae Chusik Hoeisa



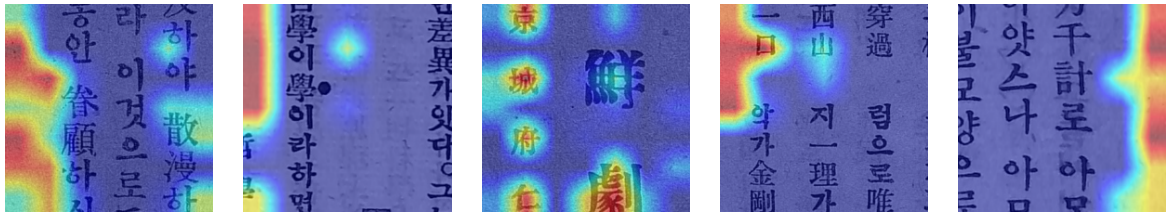
Hansŏng Toso Chusik Hoeisa



Taedong Inswaeso



Sinmungwan



Can neural networks be used to classify historical printshops and identify the specific visual features that distinguish their typographic styles?

Interpretability as a Problem

- High classification accuracy is not enough
- As humanists, we are also interested in the *why*

MIL

- *Multi Instance Learning.*[1]
- Used in the field of medical imagery.[2]
- Similar issues faced by humanists:
  - Retrieve model's decision making
  - Interpretably decision making
- We follow the *AttriMIL* implementation of Cai et al.[3]

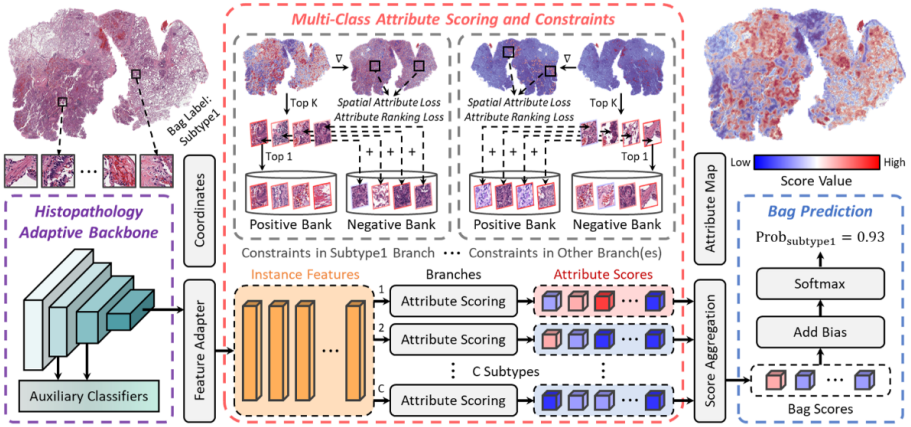
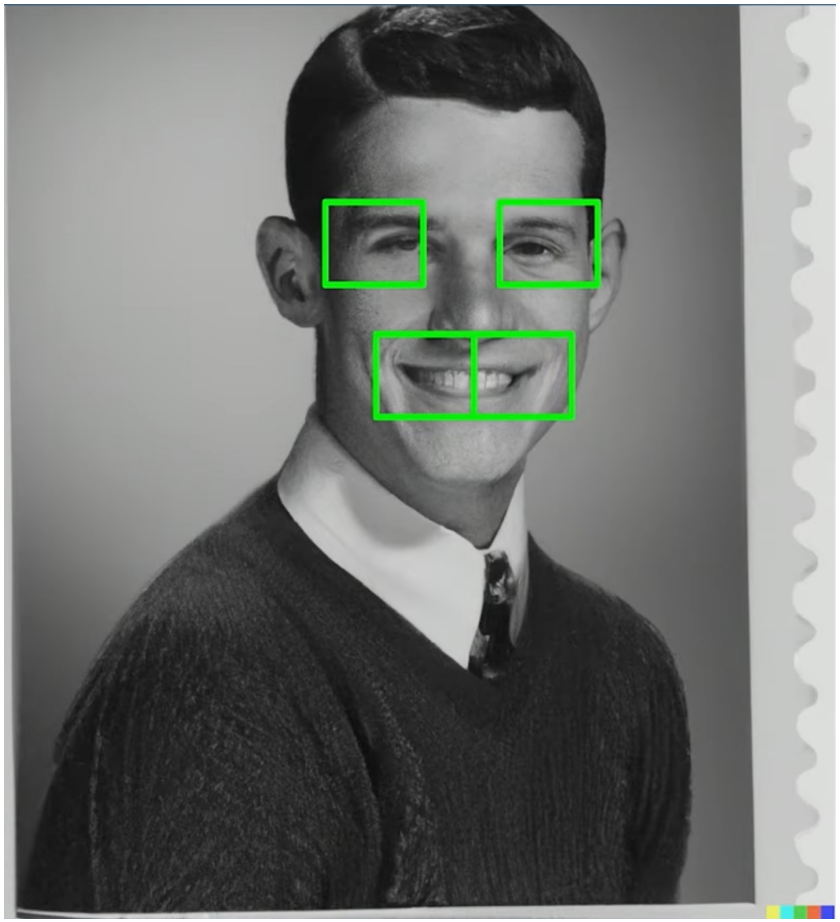


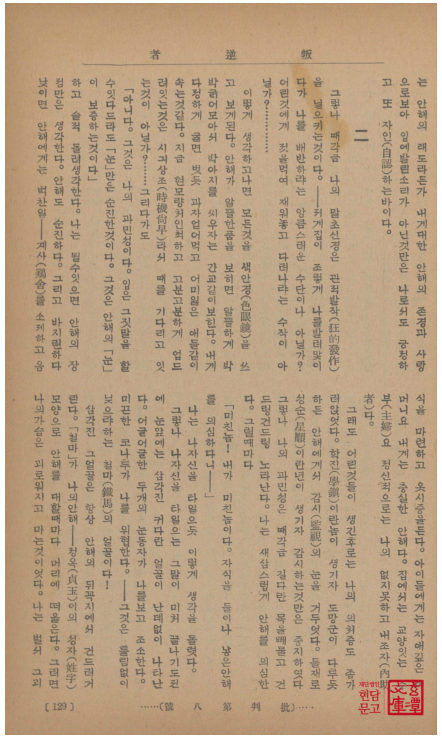
Figure 9: MIL visualization.[4]

1. Maron and Lozano-Pérez, "A Framework for Multiple-Instance Learning".
2. Papadopoulos, Topouzis, and Delopoulos, "An Interpretable Multiple-Instance Approach for the Detection of Referable Diabetic Retinopathy from Fundus Images"; Javed et al., "Additive MIL"; Yang et al., "HAMIL"; Deng et al., "Cross-Scale Multi-Instance Learning for Pathological Image Diagnosis"; Waqas et al., "Exploring Multiple Instance Learning (MIL)"; Gadermayr and Tschuchnig, "Multiple Instance Learning for Digital Pathology".
3. "Rethinking Attention-Based Multiple Instance Learning for Whole-Slide

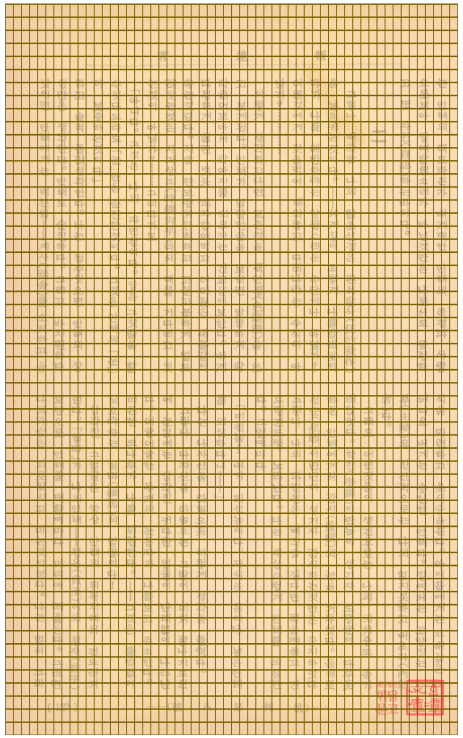
MIL Concept



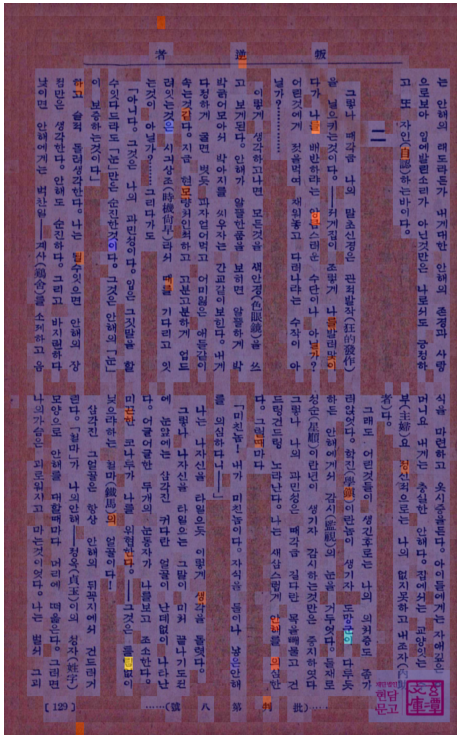
Applied



(a) Original



(b) Patched



(c) Attention displayed

Figure 10: Applied MIL

Embeddings Space

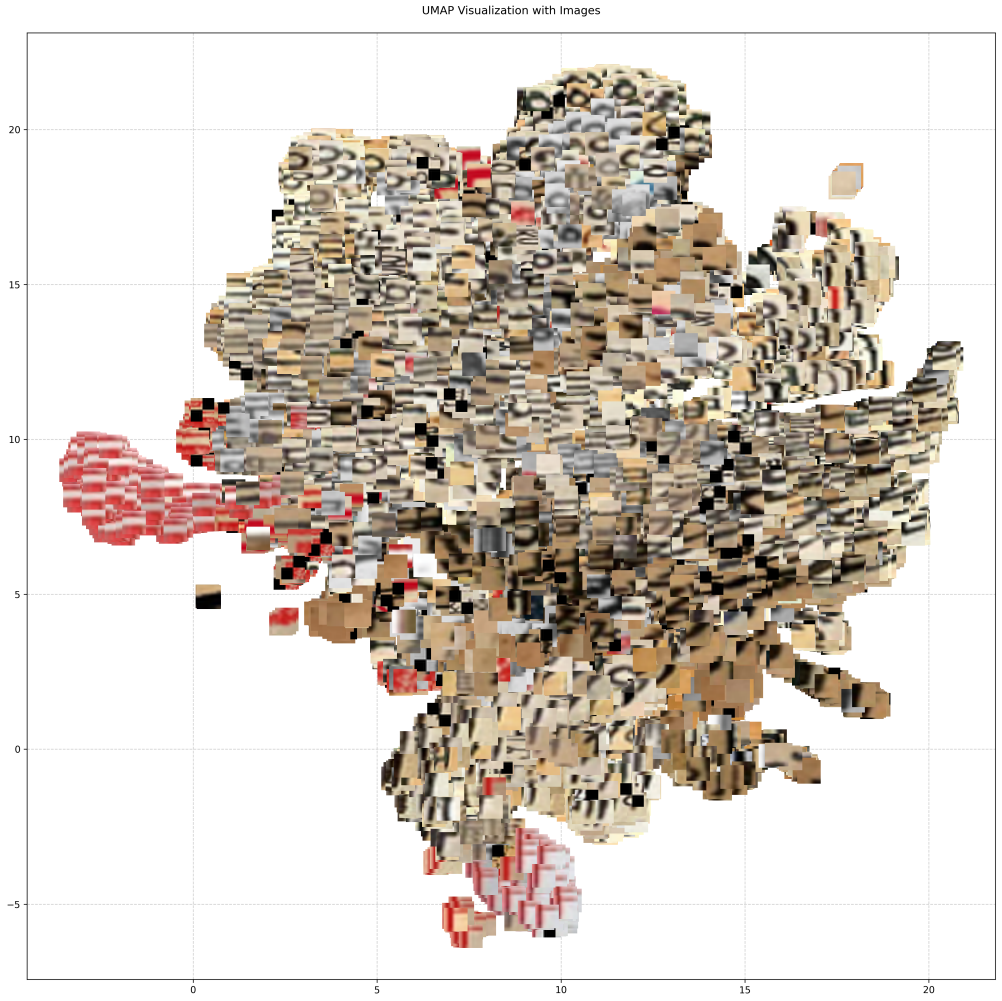


Figure 11: UMAP Visualisation of MIL Embeddings

Embeddings Space

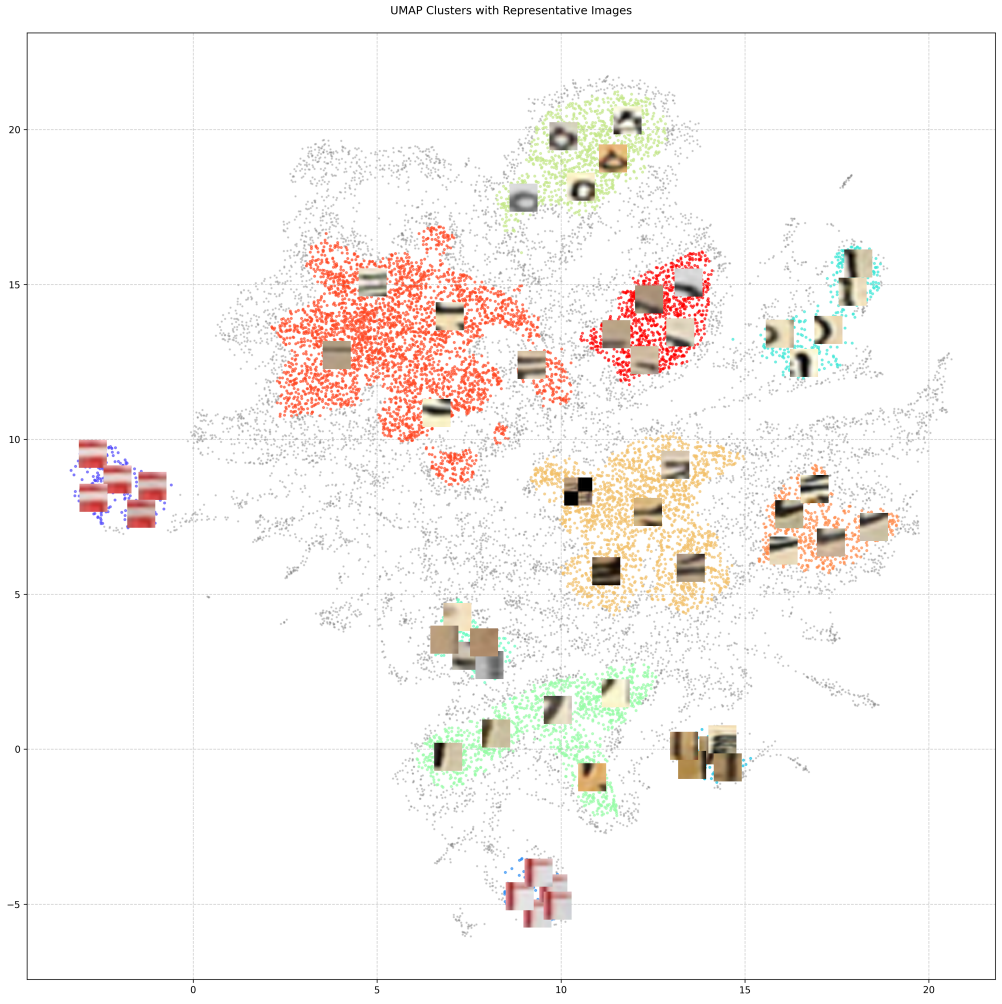
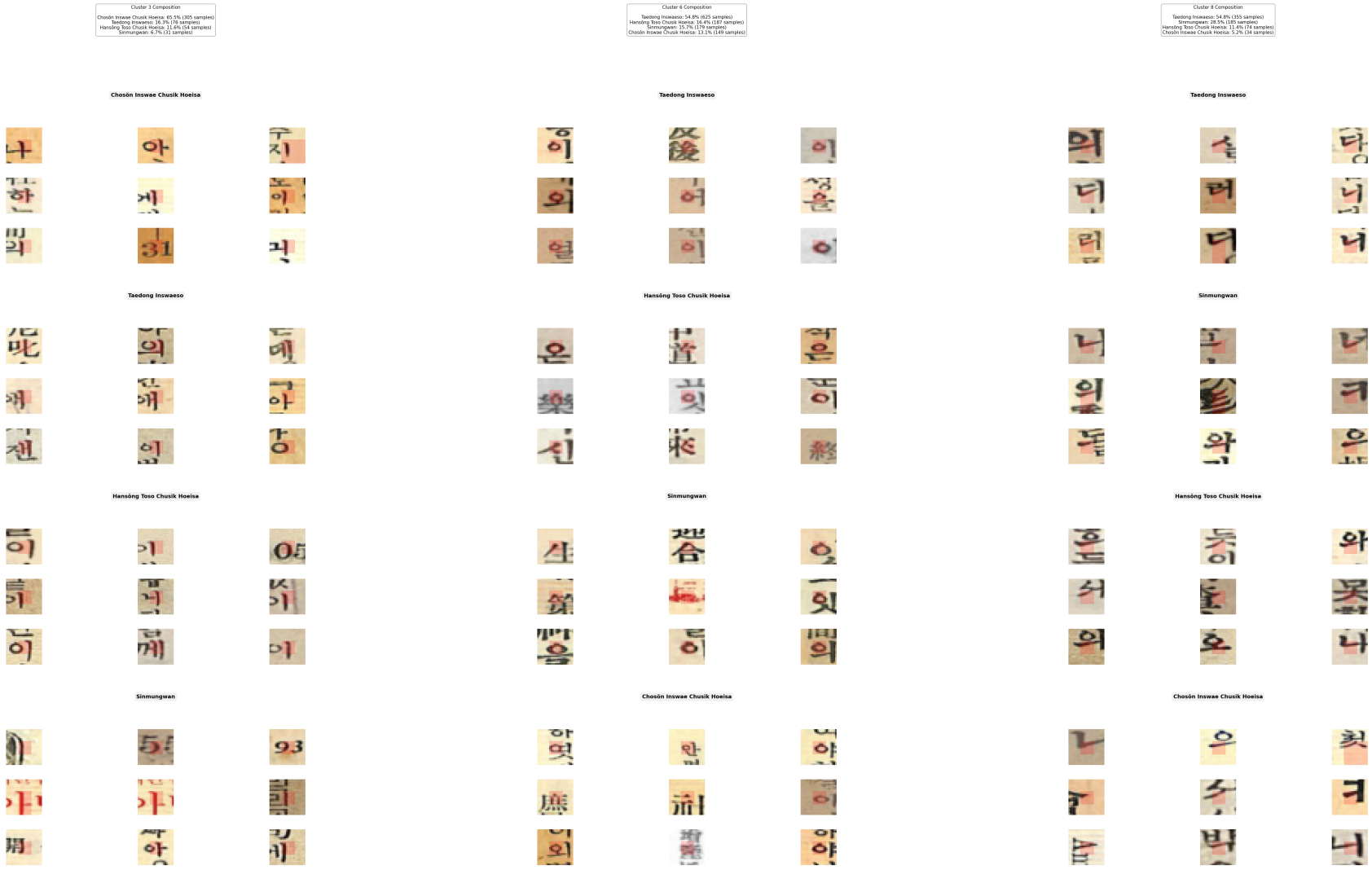


Figure 12: UMAP Visualisation of MIL Embeddings

Sampling clusters



(a) Cluster3

(b) Cluster 6

(c) Cluster 8

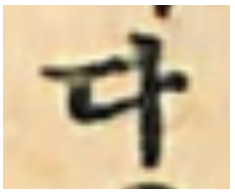
Figure 13

Compared

(a) Chosŏn Inswae Chusik Hoeisa

(b) Hansong Toso Chusik Hoeisa

(c) Taedong Inswaeso



(d) Sinmungwan

Figure 14: Sharp Stroke of ㄷ *Tigŭt* as feature for Taedong Inswaeso / Sinmungwan. Also prevalent in ㄴ *Nieun* and ㄹ *Lieul*

Features over Time

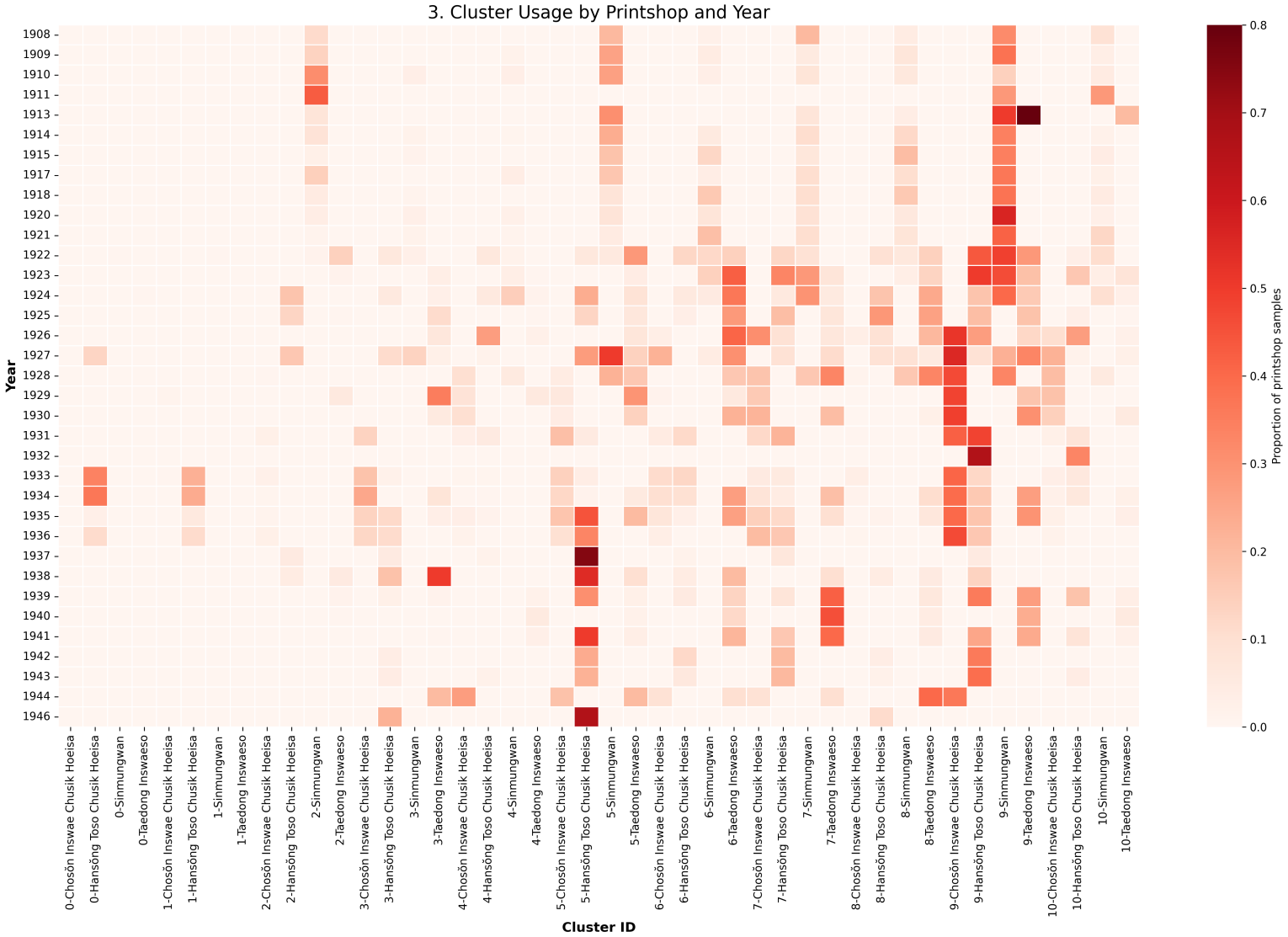


Figure 15: Heatmap of clusters division over time.

Shifts in Features

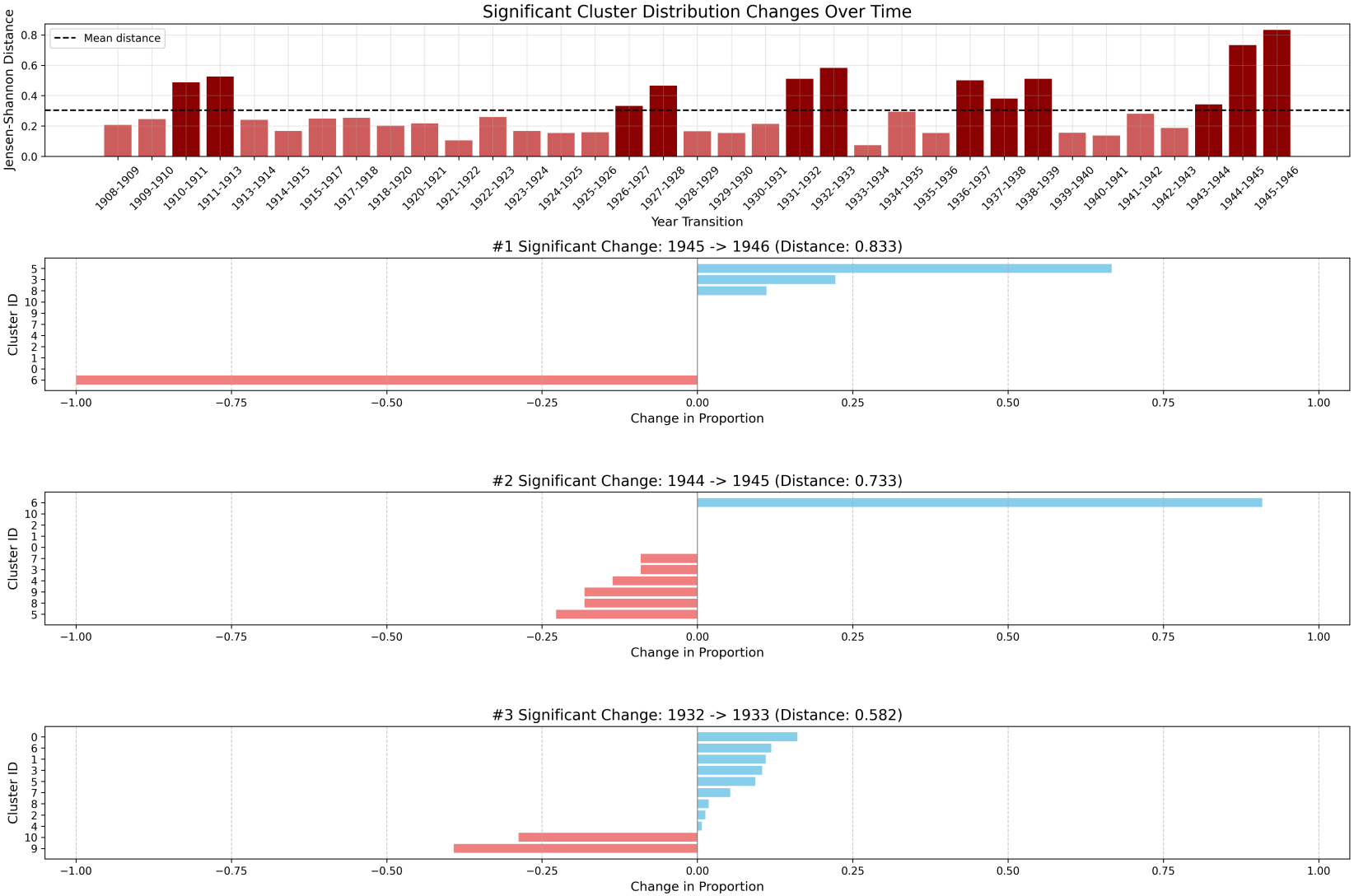


Figure 16: Cluster Shifts

Further

- Bags of Patches as feature, not singular patch
- Improvements on Clustering
- Move to typology of printshop

Cited Works

Cai, Linghan, Shenjin Huang, Ye Zhang, Jinpeng Lu, and Yongbing Zhang. "Rethinking Attention-Based Multiple Instance Learning for Whole-Slide Pathological Image Classification: An Instance Attribute Viewpoint." arXiv, March 2024. https://arxiv.org/abs/2404.00351.

Christlein, Vincent, Nikolaus Weichselbaumer, Saskia Limbach, and Mathias Seuret. "Proof of Concept: Automatic Type Recognition." arXiv, October 2020. https://arxiv.org/abs/2007.07690.

De Fremery, Peter Wayne. "How Poetry Mattered in 1920s Korea." PhD thesis, Harvard University, 2011.

Deng, Ruining, Can Cui, Lucas W. Remedios, Shunxing Bao, R. Michael Womick, Sophie Chiron, Jia Li, et al. "Cross-Scale Multi-Instance Learning for Pathological Image Diagnosis." *Medical Image Analysis* 94 (May 2024): 103124. https://doi.org/10.1016/j.media.2024.103124.

Gadermayr, Michael, and Maximilian Tschuchnig. "Multiple Instance Learning for Digital Pathology: A Review of the State-of-the-Art, Limitations & Future Potential." *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society* 112 (March 2024): 102337. https://doi.org/10.1016/j.compmedimag.2024.102337.

Hyundam Mun'go Foundation. "Hyundam Mun'go Collection." Archive, 2021.

Javed, Syed Ashar, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. "Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology." arXiv, October 2022. https://doi.org/10.48550/arXiv.2206.01794.

Lundberg, Scott M, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc., 2017.

Maron, Oded, and Tomás Lozano-Pérez. "A Framework for Multiple-Instance Learning." *Advances in Neural Information Processing Systems* 10 (1997).

Minjok Taep'yo 33-in 民族代表 33人. "Declaration of Independence," March 1919.

Papadopoulos, Alexandros, Fotis Topouzis, and Anastasios Delopoulos. "An Interpretable Multiple-Instance Approach for the Detection of Referable Diabetic Retinopathy from Fundus Images." *Scientific Reports* 11, no. 1 (July 2021): 14326. https://doi.org/10.1038/s41598-021-93632-8.

Park, Chan-seung. "3.1 Tongnip Sŏnŏnsŏ Inswae Kwajŏng kwa P'anbon ŭi Kŏmt'o 3.1독립선언서 인쇄과정과 판본의 검토 [A Review of the Printing Process and Editions of the Korean Declaration of Independence in 1919]." *Tong Asia Munhwa Yŏn'gu* 동아시아문화연구, no. 80 (January 2020): 83–114.

Schofield, Frank W. "Image of Crowd Outside of City Hall Taken by Dr. Schofield on March 1, 1919." In *The Case of Korea: A Collection of Evidence on the Japanese Domination of Korea, and on the Development of the Korean Independence Movement*, between 1919 and 1922. London: George Allen & Unwin, Ltd., 1922.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." *International Journal of Computer Vision* 128, no. 2 (February 2020): 336–59. https://doi.org/10.1007/s11263-019-01228-7.

Seuret, Mathias, Saskia Limbach, Nikolaus Weichselbaumer, Andreas Maier, and Vincent Christlein. "Dataset of Pages from Early Printed Books with Multiple Font Groups." In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, 1–6. HIP '19. New York, NY, USA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3352631.3352640.

Shin, Michael. *Korean National Identity Under Japanese Colonial Rule: Yi Gwangsu and the March First Movement of 1919*. 1st edition. Routledge, 2018.

Waqas, Muhammad, Syed Umaid Ahmed, Muhammad Atif Tahir, Jia Wu, and Rizwan Qureshi. "Exploring Multiple Instance Learning (MIL): A Brief Survey." *Expert Systems with Applications* 250 (September 2024): 123893. https://doi.org/10.1016/j.eswa.2024.123893.

Yang, Jinhyeok, Heebeom Kim, Hyobin Kwak, and Injung Kim. "HanFont: Large-Scale Adaptive Hangul Font Recognizer Using CNN and Font Clustering." *International Journal on Document Analysis and Recognition (IJDAR)* 22, no. 4 (December 2019): 407–16. https://doi.org/10.1007/s10032-019-00337-w.

Yang, Yang, Yanlun Tu, Houchao Lei, and Wei Long. "HAMIL: Hierarchical Aggregation-Based Multi-Instance Learning for Microscopy Image Classification." *Pattern Recognition* 136 (April 2023): 109245. https://doi.org/10.1016/j.patcog.2022.109245.