Scalable Optimization of Neighbor Embedding for Visualization (Supplemental Document)

Brief description of the datasets

- Iris: the UCI *Iris* dataset.
- ORL: the AT&T ORL database of face images, each image of size 92×112
- COIL: the *COIL-20* dataset from Columbia University Image Library, toy images of different angles, each image of size 128×128.
- Seg: the UCI *Image Segmentation* dataset, image patches from 7 outdoor images, originally with 19 high-level features.
- WebKB: the *WebKB4* dataset from CMU Text Learning group, text documents; 10,000 words with maximum information gain are preserved.
- 7Sectors: the *4 Universities* dataset from CMU Text Learning group, text documents classified to 7 sectors; 10,000 words with maximum information gain are preserved.
- OptDig: the UCI *optical recognition of handwritten digits*, originally with 64 dimensions.
- Reuters: the UCI *Reuters-21578* dataset, text documents, with 18933 words.
- RCV1: text documents from four classes, with 29992 words.
- Spam: A database for spam email classification, 448 numerical features for each email.
- PegDig: the UCI *pen-based recognition of handwritten digits* dataset, originally with 16 dimensions.
- Magic: the UCI *MAGIC Gamma Telescope Data Set*, 11 numerical features.
- Shuttle: the UCI *Statlog (Shuttle) Data Set*, 9 numerical features.

Table 1: Dataset statistics

Dataset	#samples	#classes	Domain	Source
Iris	150	3	biology	UCI
ORL	400	40	image	ORL
COIL	1440	20	image	COIL
Seg	2310	7	image	UCI
WebKB	4196	4	texts	CMUTE
7Sectors	4556	7	texts	CMUTE
OptDig	5620	10	image	UCI
Reuters	8293	65	texts	UCI
RCV1	9625	4	texts	RCV1
Spam	10000	2	email	SPAM
PenDig	10992	10	image	UCI
Magic	19020	2	telescope	UCI
Shuttle	58000	7	astronomy	UCI
MNIST	70000	10	image	MNIST
Covertype	581012	5	forest	UCI
TIMIT	1345233	49	speech	TIMIT

- MNIST: handwritten digit images, each of size 28×28.
- Covertype: the UCI *Covertype Data Set*, 54 numerical features.
- TIMIT: phoneme classification for speech, from the *TIMIT database*. We used 39ms time window and MFCC features. 48 semantic classes plus a miscellaneous class.

Table 2: Data sources

UCI	`http://archive.ics.uci.edu/ml/`
ORL	`http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html`
COIL	`http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`
MNIST	`http://yann.lecun.com/exdb/mnist/`
WEBKB	`http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/`
CMUTE	`http://www.cs.cmu.edu/~TextLearning/datasets.html`
RCV1	`http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/`
SPAM	`https://noppa.aalto.fi/noppa/kurssi/t-61.3050/etusivu`
TIMIT	`http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`