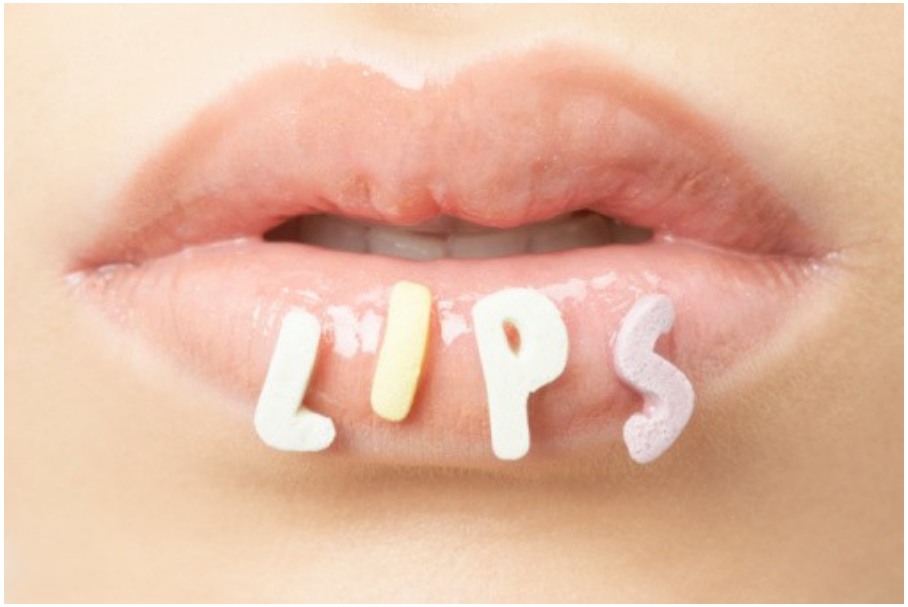
Visual Speech Recognition

Tian Zhang
Tongshuang Wu
Yi-Sheng Hsieh



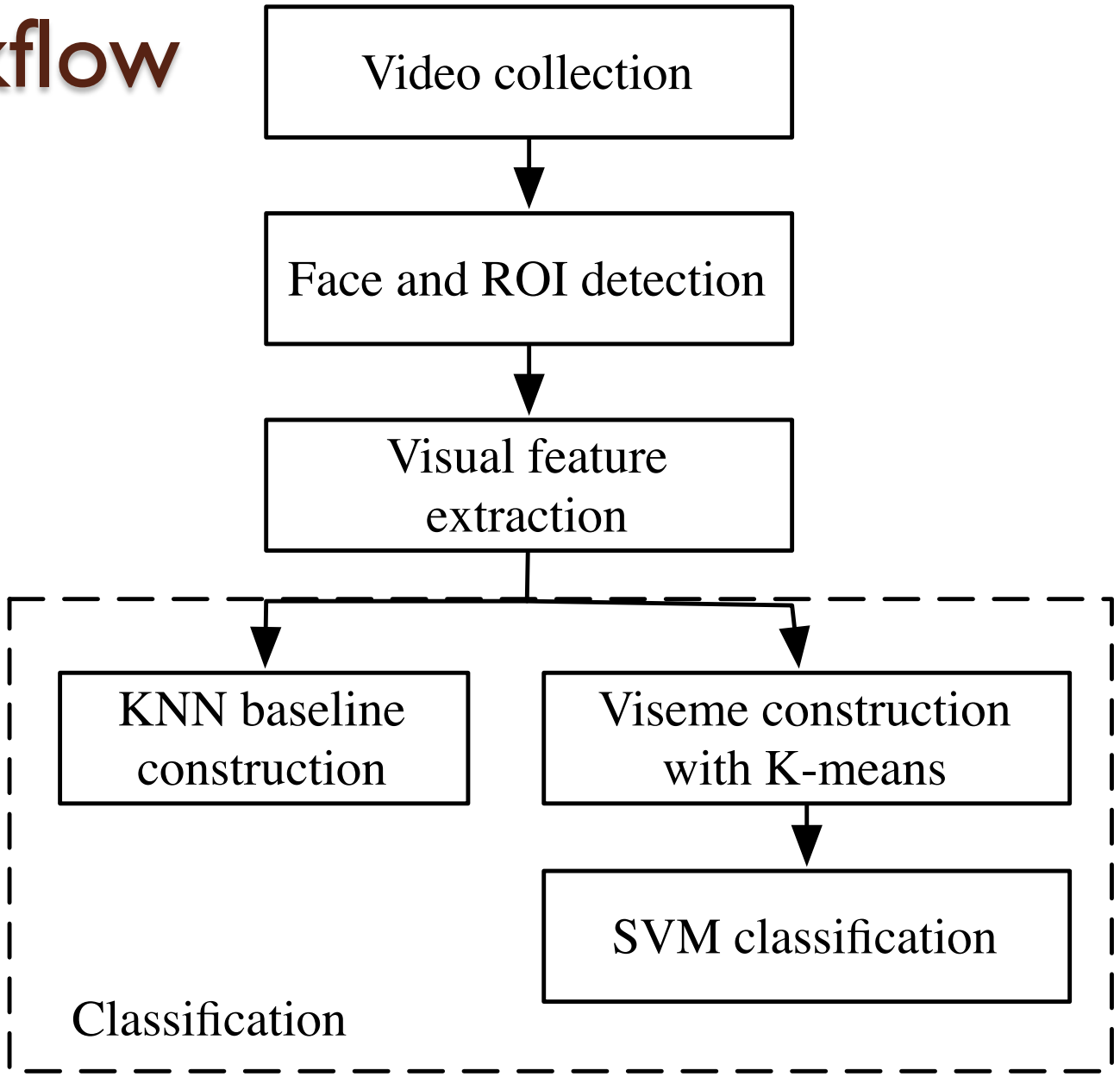
Visual Speech Recognition?

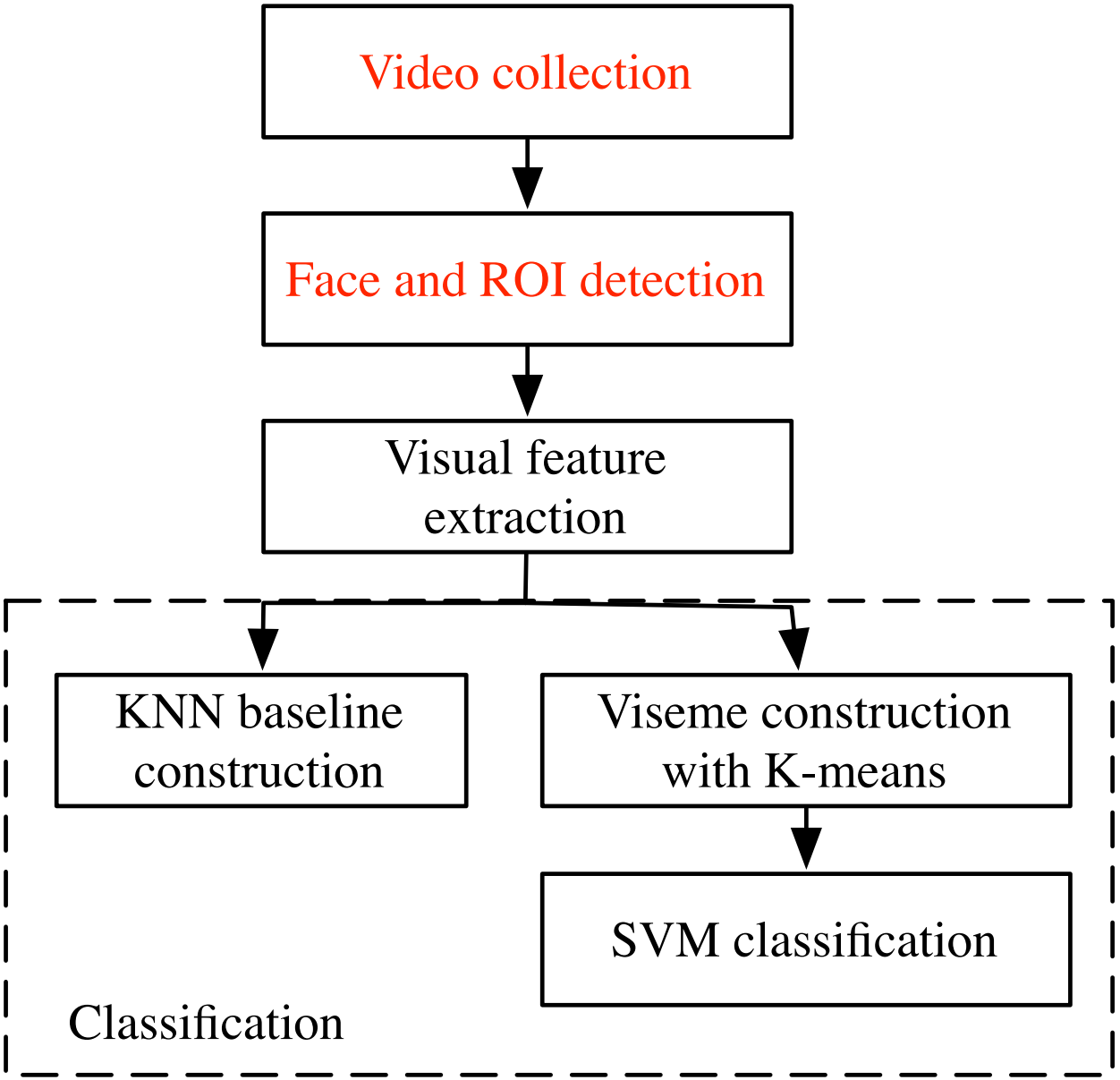
What?	lip reading by computer!!
How?	Lip motion recognition
Why?	- Useful when audio information cannot be effectively conveyed - Great social impact: for those with auditory dysfunction

What work have been done?

- Several specialized visual system for:
  - Digits
  - Chinese characters
  - Etc.
- Why it's not good enough:
  - Visual clues + automatic speech recognition
  - How about extend visual features to the full extend?

Workflow


Video collection
Face and ROI detection
Visual feature extraction
KNN baseline construction
Viseme construction with K-means
SVM classification
Classification


Video collection
Face and ROI detection
Visual feature extraction
KNN baseline construction
Viseme construction with K-means
SVM classification
Classification

Dataset

- Video source : self-made
- 2 people
- 2 word : beach dark
- 10 times

Mouth detection

- **MouthDetect = vision.CascadeObjectDetector ('Mouth','MergeThreshold',40);**
  - **Viola-jones algorithm**
  - Classification model = mouth
  - MergeThreshold

Sample result

- Word : Beach

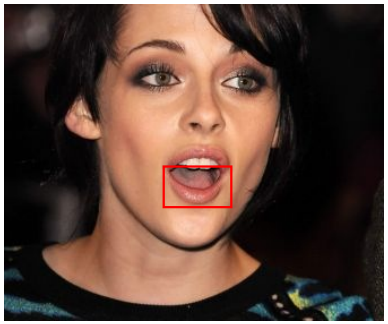
Dataset

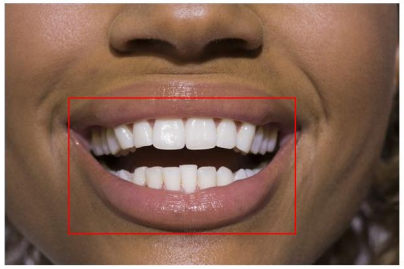
- Video source :CMU-AMP database
- 3 people
- 150 word
- 10 times

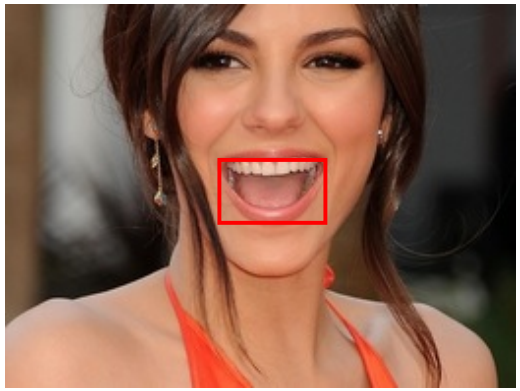
Problem

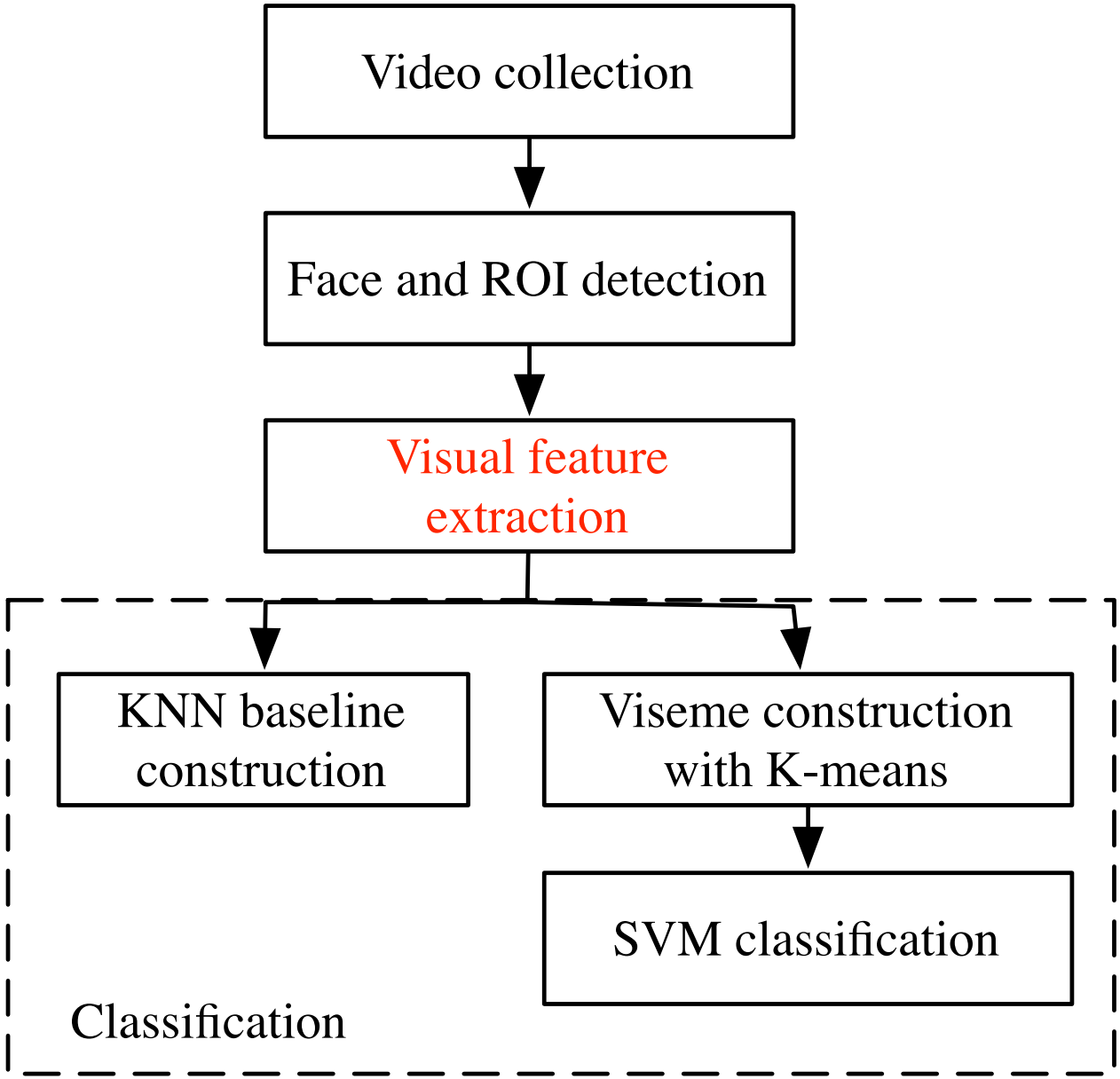
- Imperfect ROI

Observation





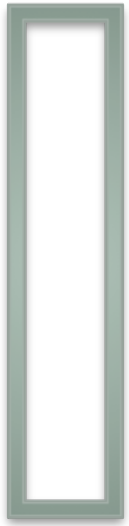



Video collection
Face and ROI detection
Visual feature extraction
KNN baseline construction
Viseme construction with K-means
SVM classification
Classification

Matrix Representation

- Width
- Length
- Mutual Information
- Quality Measure Feature
- The ratio of vertical to horizontal features
- The ratio of vertical to horizontal edges
- The proportion of tongue
- The proportion of teeth

8 x 1
Vector
Representation



Feature Description

- Width, Height
- Mutual Information: temporal correlation between two frames

$$M(X;Y)=\sum_{x}\sum_{y}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

- Quality Measure Feature: distortion between two frames

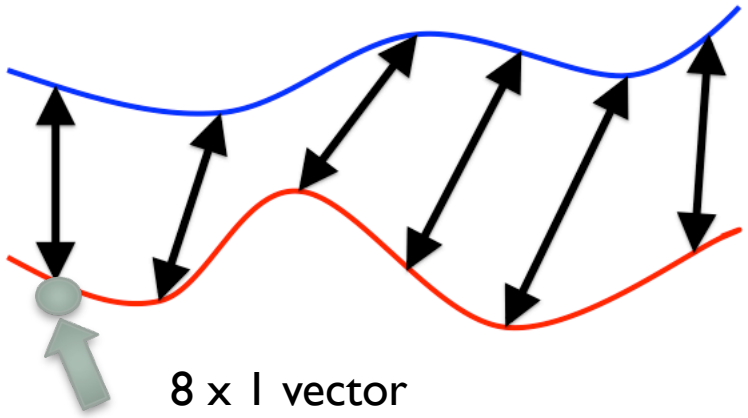
$$Q=\frac{4\sigma_{xy}\bar{x}\,\bar{y}}{(\sigma_x^2+\sigma_y^2)[(\bar{x})^2+(\bar{y})^2]}$$

- The ratio of vertical to horizontal features: High deviation from the mean in the Discrete Wavelet Transform domain is usually associated with feature.
- The ratio of vertical to horizontal edges : Sobel Edge detector

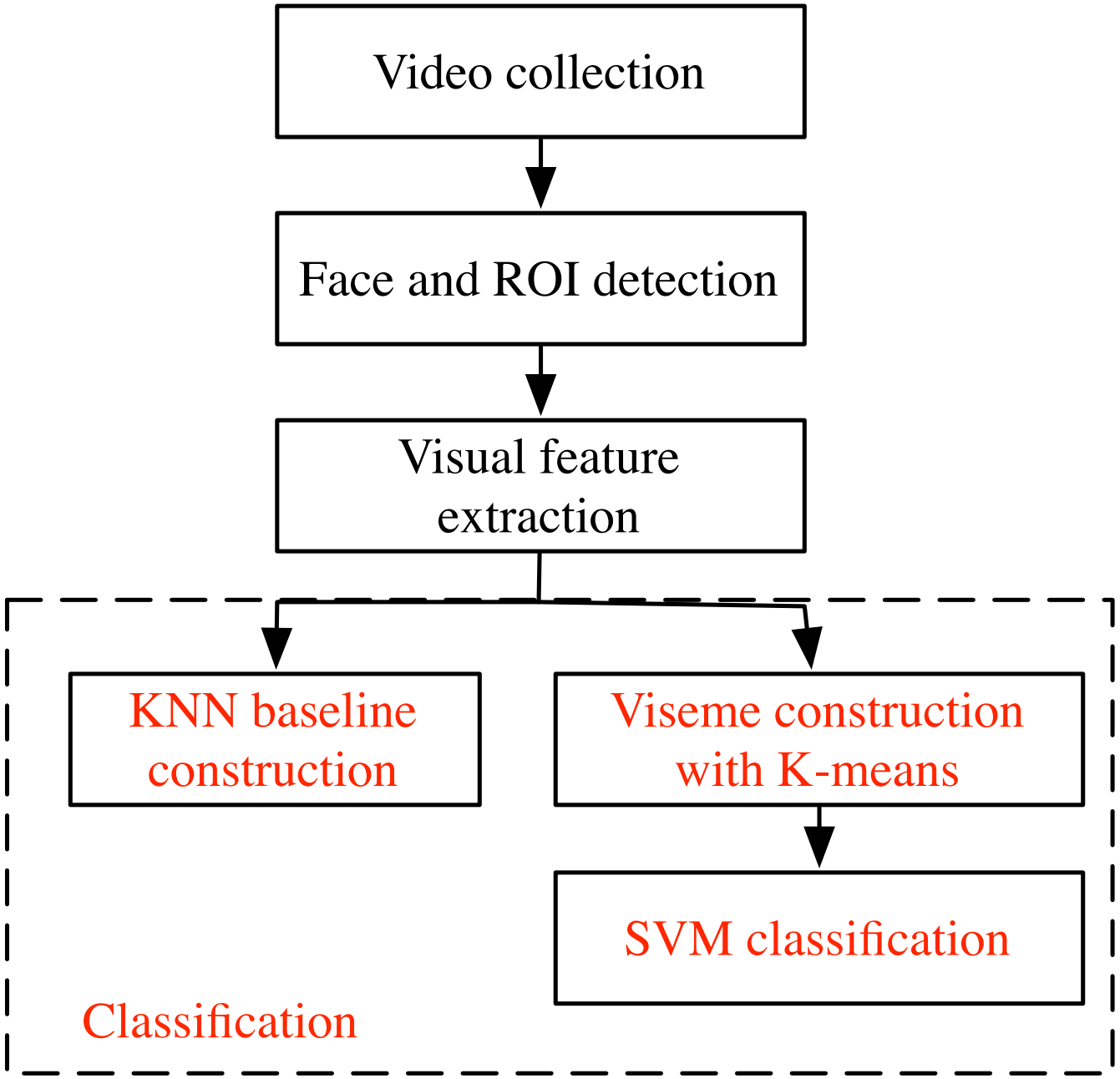
- The proportion of tongue: proportion of dark red color
- The proportion of teeth: proportion of white color

Dynamic Time Warping

- DTW: Distance between two time series of different length.



- Can be used in K Nearest Neighbor

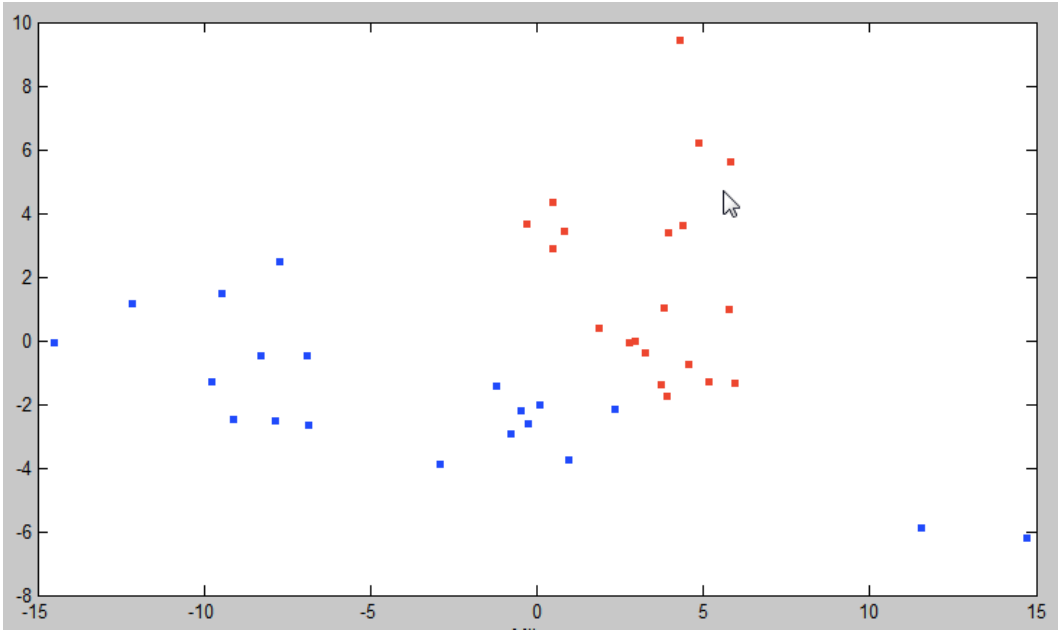

Video collection
Face and ROI detection
Visual feature extraction
KNN baseline construction
Viseme construction with K-means
SVM classification
Classification

Classification

- Most popular techniques:
  - K Nearest Neighbors (used as baseline!)
  - Hidden Markov Model (borrowed its idea for state construction!)
- What we used:
  - **HMM-like data processing, with k-means and SVM technique**

Baseline: KNN

- K-Nearest Neighbor algorithm
- Two words spoken by three people
- Accuracy = 75%



Hybrid K-means SVM

K-means on feature vectors to cluster viseme with cluster centroid, with cross validation for determining k

↓

Construct k by k feature matrix for all the samples, each entry (i, j) means the probability of transforming from state i to state j

↓

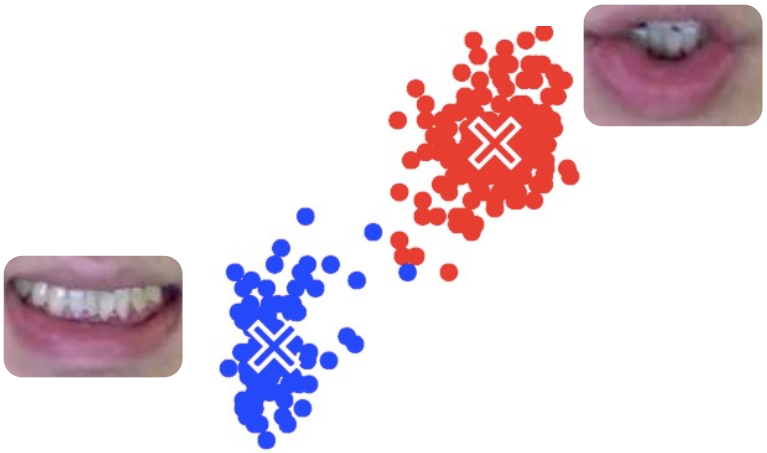
Use training data to train the SVM, with states being a cluster resulted in K-means

↓

Process test data in the same way as did to training data,

↓

Compare test and training model similarity with 1NN



1 2 3 3 1 1 3 2….

$$\begin{bmatrix} 0.1 & 0.5 & 0.09 \\ 0.11 & 0.25 & 0.01 \\ 0.08 & 0.1 & 0.31 \end{bmatrix}$$

Hybrid K-means SVM

Results comparison using provided features, full size data		
	CMU features (only used height and width)	Our features
KNN	N.A.	~75%
Kmeans-SVM	~80%	~100%

- Feature out-performs!
- Kmeans-SVM works!
- *How about multiple words?*

Guideline

- Problem description and literature review
- Technical and experimental description
  - Region of Interest detection
  - Feature extraction
  - Classification
- Future work

Future Work

- Use customized model for better ROI extraction
- Shoot some more standard videos on our own for classifying multiple word testing

Results comparison using provided features, full size data		
	CMU features	Our features
KNN	15%	?
Kmeans-SVM	?	?

Reference

- [1] G.Erten.Audiovisualspeechprocessing,2001.
- [2] C.G.Fishe, .Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, 11(4):796–804, 1968.
- [3] A. B. A. Hassanat. Visual speech recognition. *CoRR*, abs/1409.1411, 2014.
- [4] O. H. Jensen. *Implementing the Viola-Jones face detection algorithm*. PhD thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2008.
- [5] K. Kumar, T. Chen, and R. M. Stern. Profile view lip reading. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–429. IEEE, 2007.
- [6] K. Messer, J. Matas, J. Kittler, and K. Jonsson. Xm2vtsdb: The ex- tended m2vts database. In *In Second International Conference on Au- dio and Video-based Biometric Person Authentication*, pages 72–77, 1999.
- [7] A.V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Mur- phy. A coupled hmm for audio-visual speech recognition. In *Acous- tics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 2, pages II–2013. IEEE, 2002.
- [8] T.-L.Pao,W.-Y.Liao,andY.-T.Chen.Audio-visualspeechrecognition with weighted knn-based classification in mandarin database. In *In- telligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on*, volume 1, pages 39– 42. IEEE, 2007.
- [9] L. Rabiner. A tutorial on hidden markov models and selected applica- tions in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Thanks!
Any questions?