Visual Speech Recognition

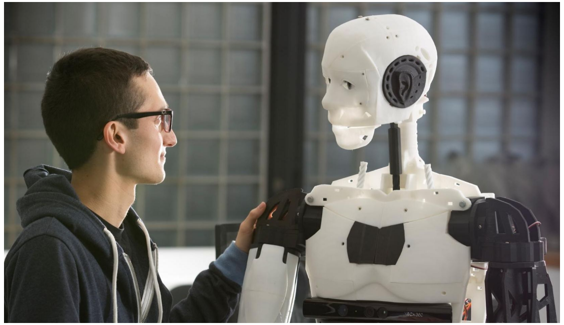
Andrea Burns

IVC Final Project

Why care about VSR?

- Inspired by lip-reading which humans use to understand language
  - Especially the hearing-impaired

- Many applications:
  - Video-text translation/generation
  - Robot instructions in noisy environments



Goal of the Project

- Perform visual speech recognition of the AVLetters Dataset*
  - 10 speakers speaking the English alphabet
  - Each letter repeated 3 times by one speaker
- Survey 4 several popular feature methods for VSR
- Compare classification accuracy to published results

*start with smaller multiclass case

Example Frames

Consonant V

Consonant M

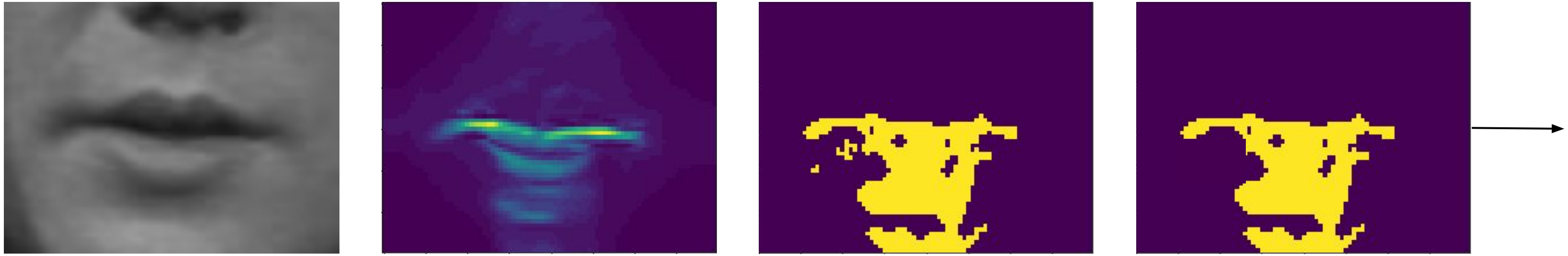
Consonant G

Ideas from Class + Literature

1) Preprocess
    a) Denoising + image differencing + SWT + binarization + erosion + artifact removal
    b) Median blur + histogram equalization
2) Obtain Features
    a) **Hu moments, Zernike moments**
        i) Yau et. al *Visual Speech Recognition Using Image Moments and Multiresolution Wavelet Images* in *IEEE* 2006
    b) **HOG descriptors**
        i) Caner Berkay Antmen, Eric Bannatyne, *Protecting the Mission: Hidden Semi-Markov Models for Visual Speech Recognition*
    c) **LBP-TOP features**
        i) Frisky et. al *Lip-Based Visual Speech Recognition System* in IEEE 2015
        ii) Guoying Zhao and Matti Pietikaeinen, *Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions*,, IEEE 2007
3) Classify
    a) 80/20 - train/test split
    b) **Support Vector Machine (SVM) OVR** classifier

Preprocessing Examples

(a)



Hu

Zernike

Denoising + image differencing + SWT   Binarization + erosion   Artifact removal

(b)

HOG

LBP-TOP

Median blur   Histogram equalization

Zernike moments

Even polynomials

$$Z_n^m(\rho,\varphi) = R_n^m(\rho)\,\cos(m\,\varphi)$$

Odd polynomials

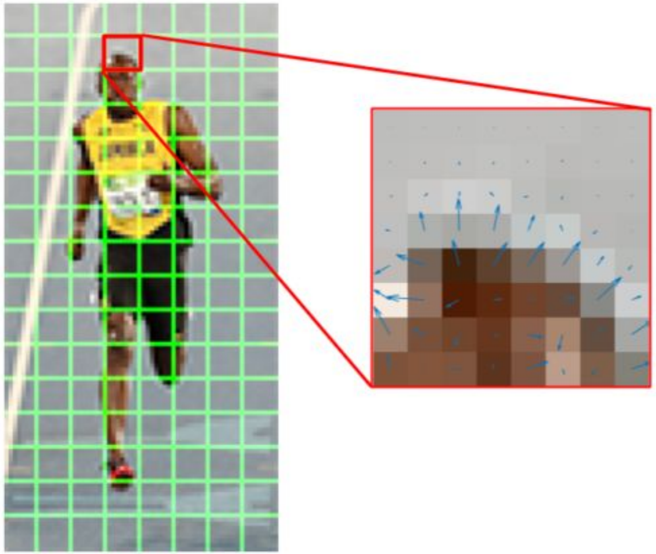
$$Z_n^{-m}(\rho,\varphi) = R_n^m(\rho)\,\sin(m\,\varphi),$$

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k\,(n-k)!}{k!\left(\frac{n+m}{2}-k\right)!\left(\frac{n-m}{2}-k\right)!}\,\rho^{n-2\,k}$$

- **Orthogonal polynomials**, no redundancy in information
- Computed up to the **9th order** based off of experimental fine tuning

HOG Descriptors

Histogram Of Gradients

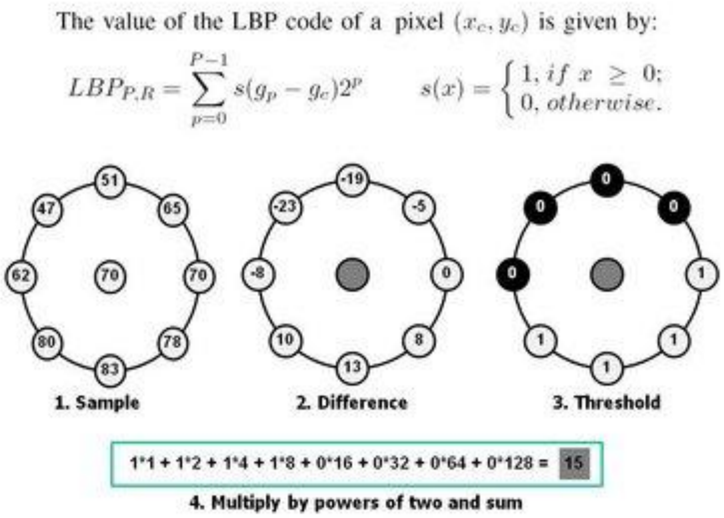


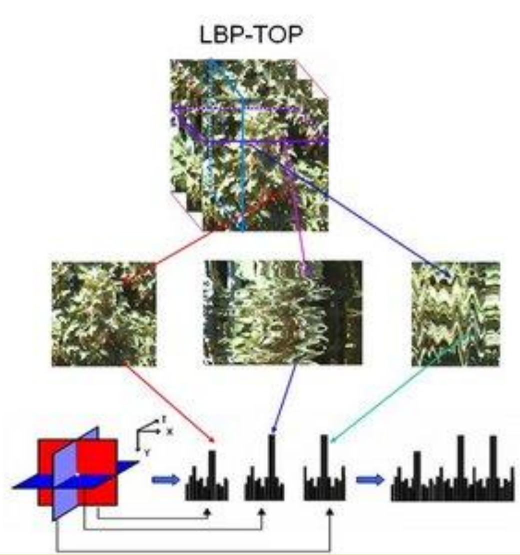
- Image is **split into smaller blocks**/cells
- Each cell has **direction gradients** (intensity changes)
- **Removes background, highlights edges**
- Histograms are **concatenated**

https://www.learnopencv.com/histogram-of-oriented-gradients/

LBP-TOP Features

Local Binary Pattern
Three Orthogonal Planes





- Obtain **texture features** over image patches in XY, XT, YT planes
- Take histograms over all patches, all frames and **concatenate XY-XT-YT**

Frame vs. Whole Video

Feature Type	Frame by Frame	Whole Video
Hu	√	**X**
Zernike	√	**X**
HOG	√	√
LBP-TOP	**X**	√

Built for speaker dependent

Built for speaker dependent, semi-independent, independent

Classification Accuracy

Feature Type	Frame by Frame 1697/425	Whole Video 72/18

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

Classification Accuracy

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

Classification Accuracy

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------

Classification Accuracy

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------
Zernike	49.85%	------

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

Classification Accuracy

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

concatenation

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------
Zernike	49.85%	------
HOG-100	68.7%	------
HOG-1000	89.17%	------
HOG-6750	**91.52%**	------

Classification Accuracy

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------
Zernike	49.85%	------
HOG-100	68.7%	------
HOG-1000	89.17%	------
HOG-6750	**91.52%**	------
HOG (10)	------	33.33%

concatenation

flattening

Classification Accuracy

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

concatenation

flattening

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------
Zernike	49.85%	------
HOG-100	68.7%	------
HOG-1000	89.17%	------
HOG-6750	**91.52%**	------
HOG (10)	------	33.33%
LBP-TOP (30)	------	**66.66%**

Classification Accuracy

*Dimensionality reduction is performed using Non-negative Matrix Factorization (NMF)

*# bins = 10 for consistency

concatenation

flattening

concatenation

Feature Type	Frame by Frame 1697/425	Whole Video 72/18
Random Guess	33.33%	33.33%
Hu	37.25%	------
Zernike	49.85%	------
HOG-100	68.7%	------
HOG-1000	89.17%	------
HOG-6750	**91.52%**	------
HOG (10)	------	33.33%
LBP-TOP (30)	------	**66.66%**
LBP-TOP-100	------	55.55%
LBP-TOP-360	------	55.55%

Findings

- Surprisingly HOG-2650 frame by frame best performance
  - LBP-TOP second + whole video representation
- Significant compression ok with HOG frame by frame
  - not the same for video LBP-TOP
- Trade-off between more samples, lower quality features + fewer samples, higher quality features
- Look into data splits
  - Speaker dependent
  - Speaker semi-dependent
  - Speaker independent

Thank you!