Using Regression Techniques to Predict Weather Signals from Image Sequences

Richard Speyer

Master of Science Thesis Defense
April 22nd, 2009

with slides from Nathan Jacobs and Robert Pless

Overview

- Introduction
- Background Information
  - AMOS Dataset
  - Weather Data
  - Related Work
- Methods & Algorithms
  - Canonical Correlation Analysis (CCA)
  - Principal Coefficient Analysis (PCA)
  - Putting everything together
- Results & Analysis
  - Wind Velocity
  - Vapor Pressure
  - Training set analysis
- Conclusion
- Future Work

Introduction

- Given a set of images collected from a static webcam and the associated ground truth weather data, can we build a model to predict the weather at a given time by just looking at the associated image?
- If so, why should we bother?
  - Help us gain a better understanding of local weather patterns
  - Fill in missing data from weather stations
  - Uses cheap sensors already in place as opposed to setting up an expensive weather station
- We will use regression and correlation techniques such as Canonical Correlation Analysis (CCA) to achieve this goal

Background Information & Related Work

Images & Webcam Data

- Images inherently contain large amounts of information beyond just pixel colors
- Our job as researchers is to develop methods to extract this information to gain a better understanding what is going on in a given scene
- Thousands of webcams are already installed all over the world and are constantly collecting data
- Provides a vast resource of "real world" data for us to test our methods

AMOS Dataset

- *The Archive of Many Outdoor Scenes (AMOS)*
  - Images from ~1000 static webcams,
  - Every 30 minutes since March 2006.
  - *www.cs.wustl.edu/amos*
- Capture variations from fixed cameras
  - Due to lighting (time of day), and
  - Seasonal and weather variations (over a year).
- Collected from cameras mostly in the USA (a few elsewhere)

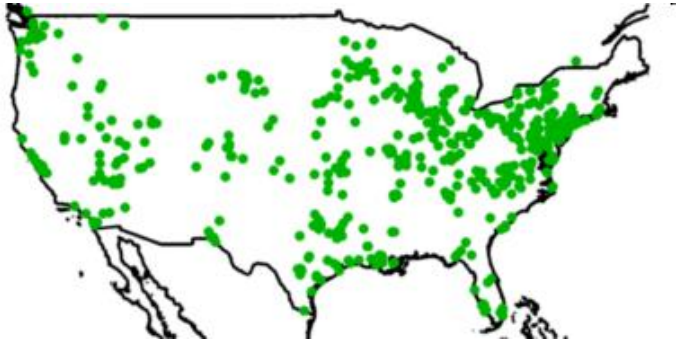
AMOS Dataset

1000 webcams

x 3 years

30 million images

Variations over a year and over a day



Weather Data

- Historical Weather Data Archives (HWDA) maintained by the National Oceanic and Atmospheric Administration (NOAA)
- Maintains hourly weather data from about 6,000 weather stations across the country from January 1, 1933 to present
- But, there are segments of missing data due to broken equipment, failure to report, etc.

Related Work

- Inferring information from large webcam datasets
  - S.G. Narasimhan, C. Wang, and S.K. Nayar. *All the Images of an Outdoor Scene*.
  - James Hays and Alexei A. Efros. *im2gps: estimating geographic information from a single image*.
  - Jean-Francois Lalonde, Srinivasa G. Narasimhan, and Alexei A. Efros. *What does the sky tell us about the camera?*
- CCA and multimedia
  - Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. *Multimedia content processing through cross-modal association*.

Methods & Algorithms

Canonical Correlation Analysis (CCA)

- Measures the linear relationship between two multi-dimensional variables
- Given two data matrices X and Y finds two matrices A and B to maximize the correlation:

$$\rho = corr(AX, BY)$$

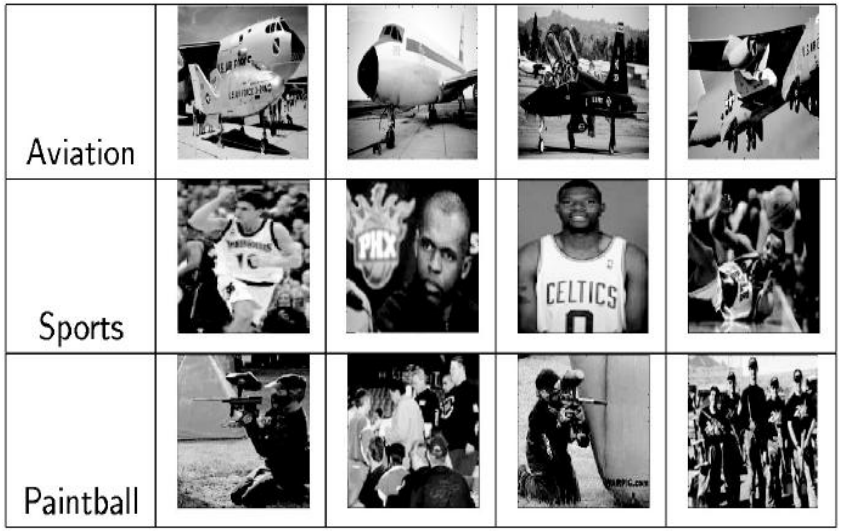
- Finds two sets of basis vectors such that the correlation between the projections of the variables onto these basis vectors is maximized
- Iteratively finds vectors which maximize the correlation between the two datasets and minimize the correlation between previously measured vectors
  - Vectors are not necessarily orthogonal

Canonical Correlation Analysis (CCA)

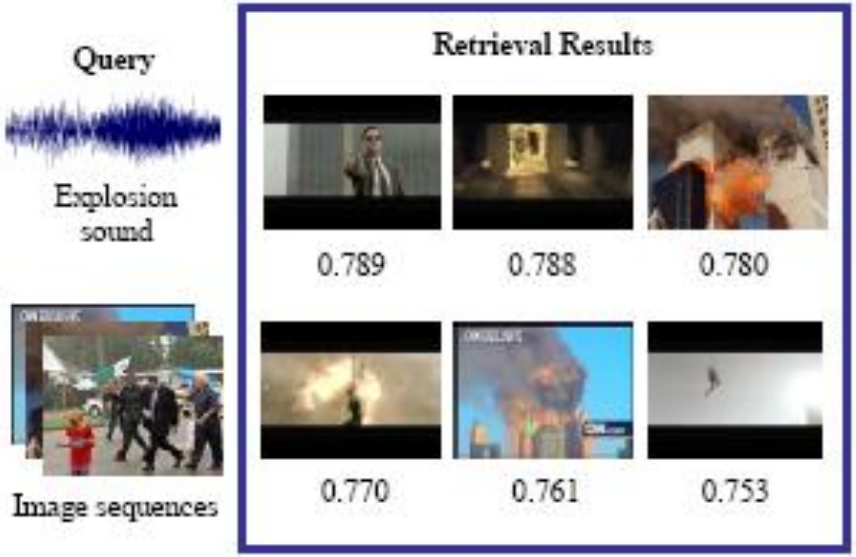
- Not dependent on the coordinate system of the variables
  - Can find correlations regardless of measurement system used, which is important since images and weather data use different forms of measurement
- Invariant to affine transformations of the variables
- It is important that each data entry in X has a corresponding entry in Y, and vice versa (same number of rows)

Canonical Correlation Analysis (CCA)

- One common application of CCA in multimedia is content-based image and video retrieval



Text-based retrieval



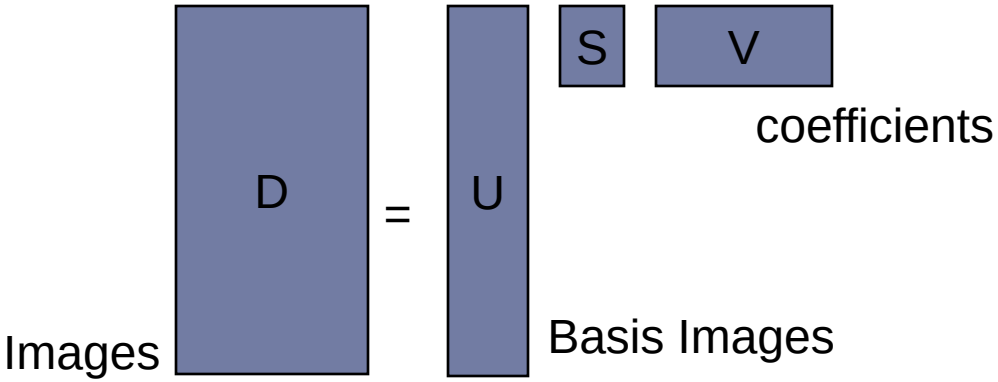
Sound-based retrieval

Principal Coefficient Analysis (PCA)

- PCA is a method used to extract the most significant features from given dataset
- Given a set of images I and a number k>0, finds the k most important features in the set of images
- [U S V] = PCA(I,k)
  - U contains the k feature, or basis, images, all of which are orthogonal to each other
  - S is a diagonal matrix which contains the weights of each feature vector
  - V contains the coefficients of each basis image for each actual image
- Will extract the most significant features to minimize

$$\sum_{i=1}^{n} (I_i - USv_i)^2$$

Principal Coefficient Analysis

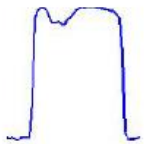
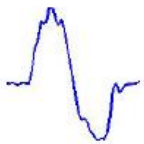
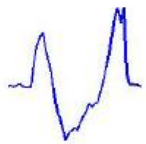
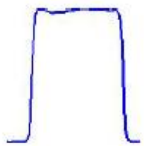
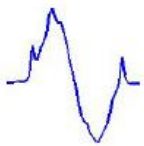
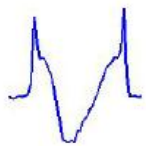
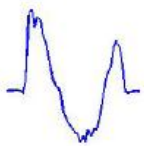
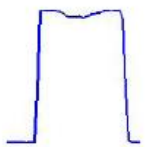
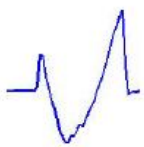


- We can reconstruct image x as a linear combination of the basis images: $i_x=USv_x$
- Reconstructed images will not exactly match the original images
  - Similarity increases as we increase the number of coefficients k



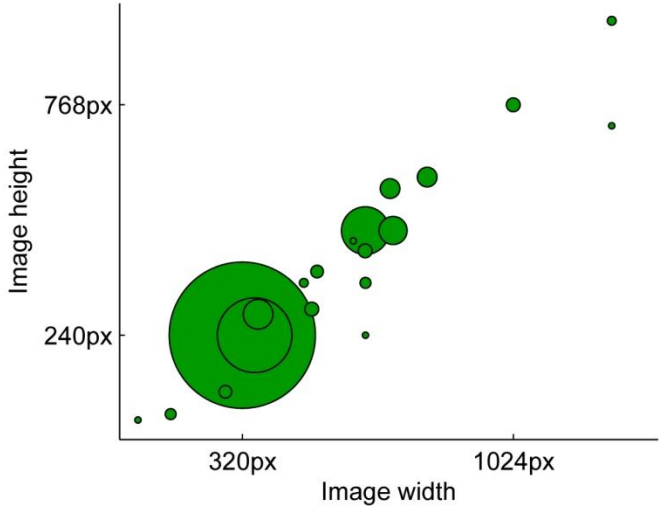

= + $f_1(t)$ + $f_2(t)$ + ...

mean Image component 1 component 2

	Mean Image	Coef 1	Coef 2	Coef 3
Camera 1				
Camera 2				
Camera 3				
Camera 4				

Why use PCA components?

- Most of the images in the AMOS dataset are 320 x 240 pixels = 76,800 pixels
- Lots of dimensions for CCA, will greatly hurt the runtime of our algorithms
- We can reduce this number dramatically by instead using the PCA components of each image (k=10)



Bringing it together - Algorithms

- Given set of timestamped images $I = i_1,\ldots,i_x$ and weather data entries $W = w_1,\ldots,w_y$
- Concurrently iterate through both sets and remove entries which do not have corresponding entries in opposite set. Both sets now have n entries
- Run PCA on images; $V = v_1,\ldots,v_n$ are vectors of length k containing the principal coefficients for each image, $U = u_1,\ldots u_k$ are the basis images
- Run CCA using V and W as input matrices
  - [A B] = CCA(V, W)

Predicting the Weather

- Given a new image i and CCA projection matrices A and B, we can predict the value of the weather signal in the following way:
  - Find the PCA coefficients of the new image I on our existing basis vectors U (**v**)
  - Value of weather signal **w** is:

$$\mathbf{w} = A\mathbf{v}B^{-1}$$

Results & Analysis

Results & Analysis

- Predict two signals which present unique challenges
  - Wind Velocity
  - Vapor Pressure
- Analyze effect of training set size on performance
- Find minimum amount of data needed to build a good predictor

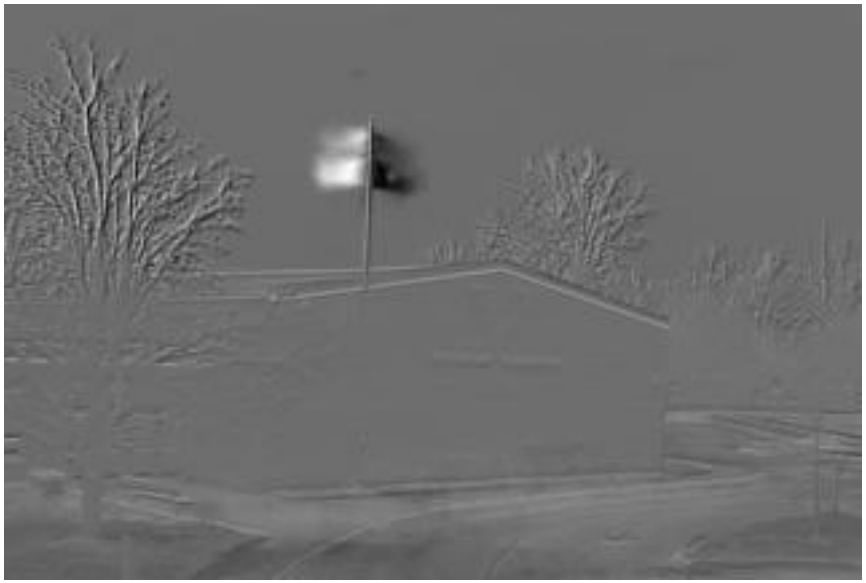
Wind Velocity

- Wind velocity is made up of 2 components: north/south and east/west
- Effects are only noticeable on very local regions of the image
  - Trees
  - Flags
- In order to ignore changes due to time of day, all images are captured between 10 AM and 2 PM
- 204 training images, 102 test images

(Left). A sample image from AMOS camera #194 in Decatur, IN

Wind Velocity

- Can visualize the derived CCA bases by projecting back onto the image space using the basis images U (U A)
- Analysis of CCA basis images shows only one accurate dimension
- Makes sense since webcam captures in 2D so we can only see the flag blowing along one direction



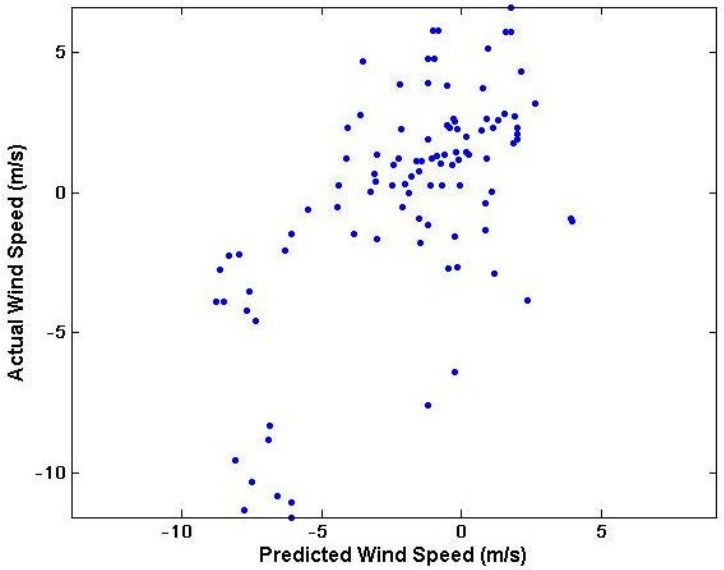
First CCA basis vector projected onto the image space

Second CCA basis vector projected onto the image space

Wind Velocity

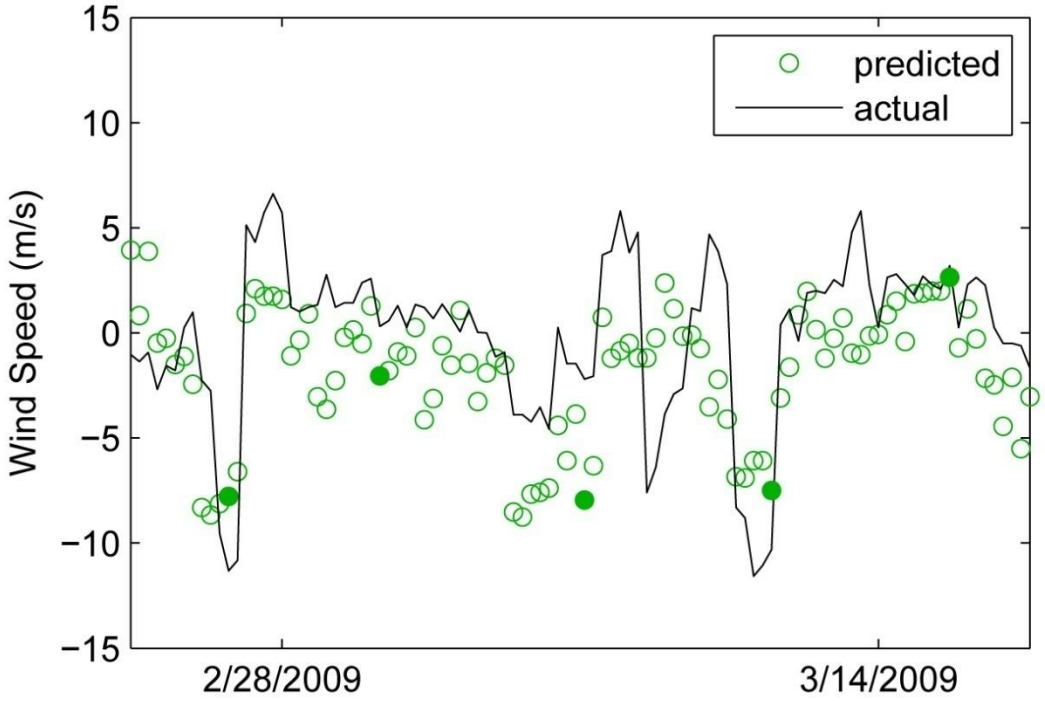
- If we sort images by the value of the first CCA coefficient, it is clear that the direction of the flag is a key feature

Sample images sorted by the value of the first CCA coefficient



Predicted vs. actual wind speeds (r=0.61759)

Wind Velocity

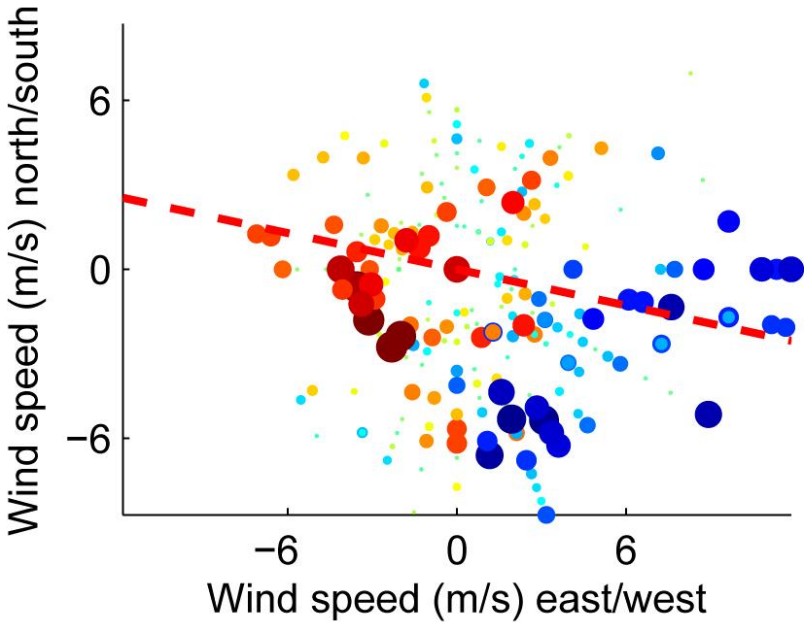


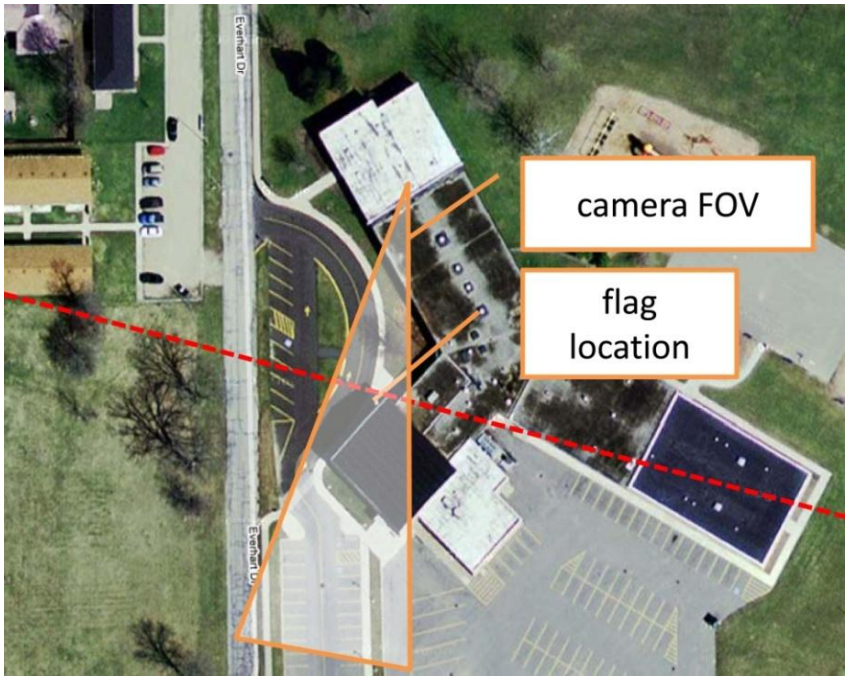
Wind Velocity

- Since flag can only be viewed in 2D, we only predicted the magnitude along some unknown vector
- Use this information to compute this unknown vector, which will tell us the orientation of the camera
- Let A = n×2 matrix containing actual wind velocity, let b = n×1 vector containing predicted wind magnitudes
- By solving Ax = b for x, we will find the orientation vector of the camera

Wind Velocity

- On the left, the size and color of each marker is determined by the predicted wind speed and the location of each marker is determined by the north/south and east/west components of the actual wind speed at the same time. The dashed red line is the normal to the projection axis determined by running linear regression between the predicted and actual values.
- On the right, we show this axis overlaid on a Google Maps image with the field of view crudely estimated by hand.





Vapor Pressure

- Contribution of water vapor to overall atmospheric pressure (millibars)
- As opposed to wind velocity, vapor pressure will likely have a more universal effect on the image
- Expect to see the cloud cover increase as vapor pressure increases
- Images captured between 10 AM and 2 PM
- 198 training images, 99 test images

(Left). A sample image from AMOS camera #619 in Houston, TX

Vapor Pressure

- CCA projection image when original images are used (left) is not very compelling. Seems to be identifying sunlight as a key indicator of vapor pressure
- When gradient images are used (right), the image is far more compelling

$$I_g = \sqrt{\frac{dI}{dx}^2 + \frac{dI}{dy}^2}$$

- Appears to be identifying the clarity/visibility of building outlines as an important indicator of vapor pressure

CCA basis vector projected onto the image space (original images)

CCA basis vector projected onto the image space (gradient images)

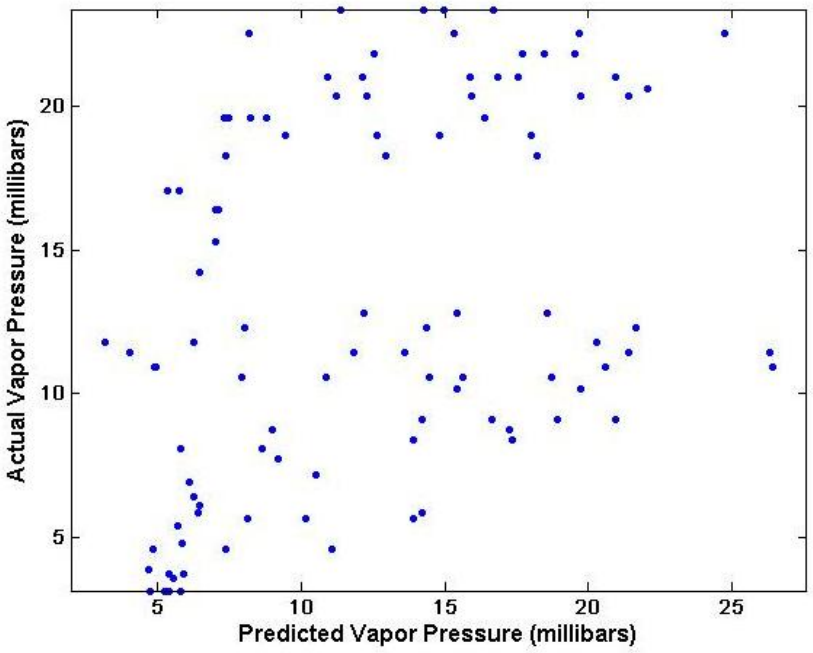
Vapor Pressure

- If we sort images by the value of the CCA coefficients, it is clear that the cloud cover and resulting visibility of the buildings is a key feature

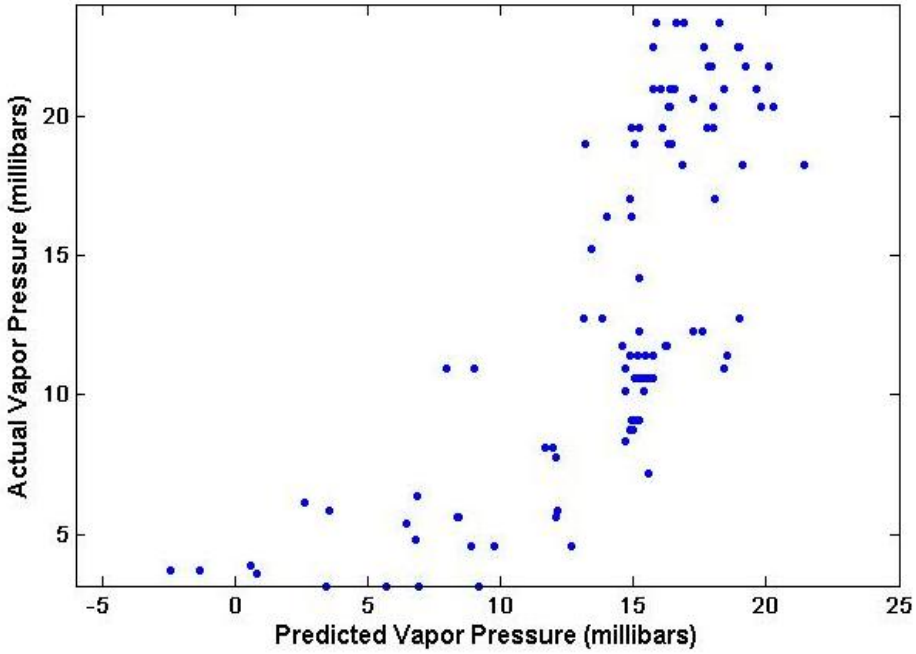
Sample images sorted by the value of their CCA coefficient

Vapor Pressure

- Scatter plots of predicted vs. actual vapor pressures (millibars) further verify the improvement with gradient images
- Agrees with expectations from CCA projection images

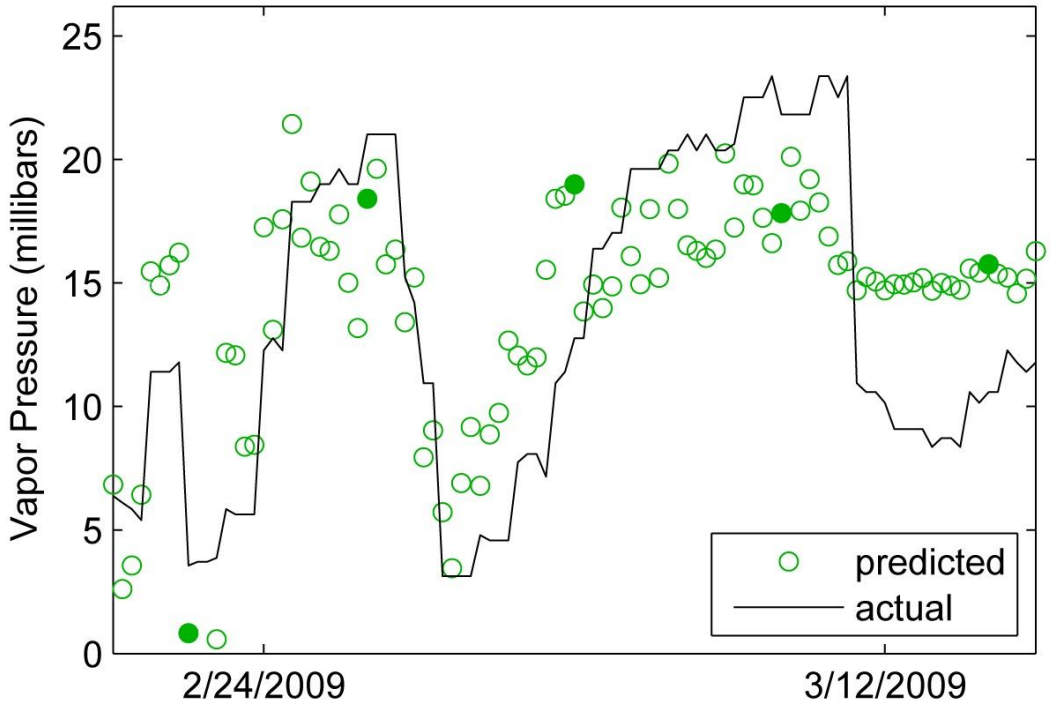


Original images (r=0.3894)



Gradient images (r=0.73684)

Vapor Pressure



Training Set Analysis

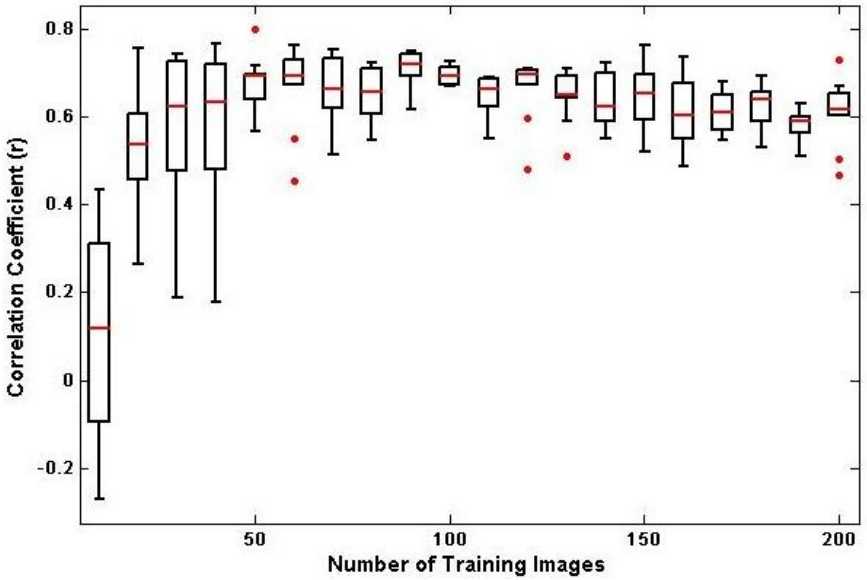
- How many images does it take to build a good weather predictor?
- Want to find out what the minimum amount of data is required to build an accurate and precise predictor of a given weather signal
- Previous studies (Barcikowski & Stevens, 1975) indicates that CCA requires between 40 and 60 times the number of variables in training data
- Might less if data variations are limited and samples sufficiently cover possible range

Training Set Analysis

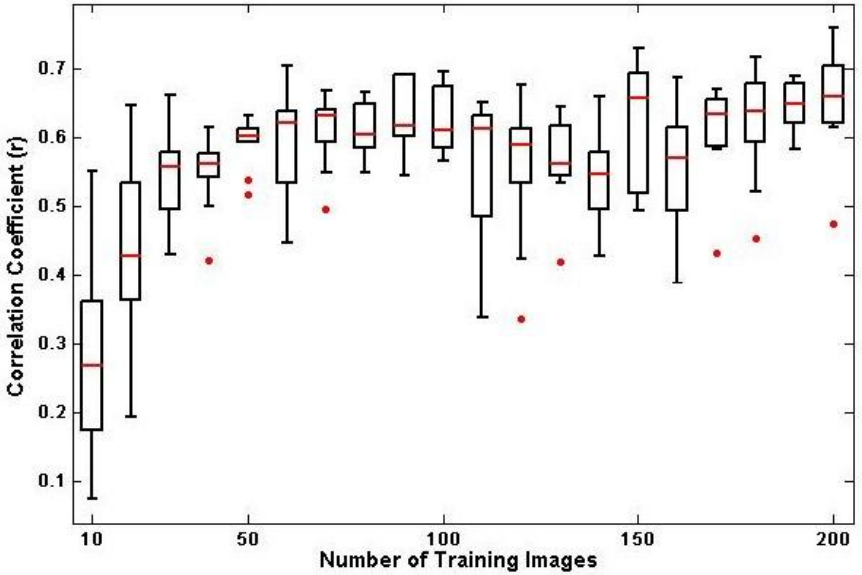
- Vary training size between 10 and 200 images
- Run PCA, CCA on images
- Take new set of 100 test images, get PCA components on existing basis vectors U
- Predict value of weather signal, compute correlation coefficient r with known actual values
- Repeat 10 times at each training size with different training sets
  - Reduces effect of lucky/unlucky selections of training data

Training Set Analysis

- Average correlation coefficient (red lines) steadily increase as training set size increases
- Stop rising rapidly around 80 images, amount of variation in coefficients decreases as well
- 80 images = 2 weeks of data



Wind Velocity (m/s)



Vapor Pressure (millibars)

Conclusion & Future Work

Conclusion

- Images carry with them lots of higher level information
- Develop methods to extract this information to gain a better understanding of what is going on at a given location
- Using regression and correlation techniques, we can predict weather signals simply by observing a given webcam
- Only 2 weeks of data is required to build an accurate model

Future Work

- Take greater advantage of AMOS dataset
- Further automate the methods used in order to apply to many cameras
- Begin to gain a better understanding of local weather variations by combining predicted weather data with locations of cameras
- Build models to work at all times of day

Acknowledgements

- Dr. Robert Pless
- Nathan Jacobs
- Thesis Examination Committee
  - Dr. Ron Cytron
  - Dr. Tao Ju
  - Dr. Robert Pless
- Media & Machines Lab

Questions?