WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Computer Science and Engineering

Thesis Examination Committee:
Dr. Ron Cytron
Dr. Robert Pless
Dr. William Smart

USING REGRESSION TECHNIQUES TO PREDICT WEATHER SIGNALS FROM IMAGE SEQUENCES

by

Richard Speyer

A thesis presented to the School of Engineering
of Washington University in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

May 2009
Saint Louis, Missouri

ABSTRACT OF THE THESIS

Using Regression Techniques to Predict Weather Signals from Image Sequences

by

Richard Speyer

Master of Science in Computer Science

Washington University in St. Louis, 2009

Research Advisor: Dr. Robert Pless

Webcams are cheap sensors that capture a potentially large amount of information about a scene. This thesis considers the use of regression and correlation techniques such as Canonical Correlation Analysis (CCA) to convert these webcams into environmental sensors and predict the values of weather signals. Local environmental properties often directly affect the images we collect from the webcams; whether it is cloudy or sunny is visible by the presence of shadows; wind speed and direction is visible in smoke, flags, or close up views of trees; particulate density is reflected in haziness and the color spectrum during sunset. Using the AMOS database, which has been archiving nearly 1,000 webcams every 30 minutes for the last 3 years, we explore relationships between the amount of training data and the accuracy with which we are able to infer the values of certain weather signals including wind speed & direction and vapor pressure from inherent properties in the image. This allows the webcams *already* installed across the earth to act as generic sensors to improve our understanding of local weather patterns and variations.

Acknowledgments

First and foremost, I would like to thank Dr. Robert Pless, who has been my research and faculty adviser for the last four years. His constant attention and guidance have been crucial in the construction of this thesis and I am extremely grateful for all of his help. My experience working under him in the Media & Machines Lab is one of the most valuable of my college career and has taught me skills that I will surely use after I graduate.

I would also like to thank Nathan Jacobs and many other students who helped review this thesis and the related research as well as Dr. William Smart and Dr. Ron Cytron for sitting on my Thesis Committee along with Dr. Pless.

Finally, I would like to thank the Department of Computer Science & Engineering for promoting and providing many opportunities for undergraduates to get involved in research projects. I have benefited greatly from working in a research lab during my time here and that would not have been possible without the department's commitment to undergraduate research.

Richard Speyer

Washington University in Saint Louis
May 2009

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The appearance of a scene can often give us an abundance of information about the scene including time of day, location, and weather. With all of this information available, it is up to us to find ways to automatically extract it from images of the scene to allow us to better understand what is going on at a given location. In this thesis we will attempt to use the information collected by webcams to infer information about environmental signals such as wind speed & direction and vapor pressure.

The challenge is that images may vary more due to other causes such as time of day than those which relate to our signals of interest. Thus, we begin by exploring mechanisms for learning features that are invariant to other scene changes. Through the use of correlation techniques such as Canonical Correlation Analysis (CCA), we find that we are able to extract useful information from webcam images which allow us to predict local weather data and as a result gain better understanding of local weather patterns and variations. Furthermore, this allows us to use the abundantly existing webcams all over the country as crude weather sensors instead of depending only on the government weather stations sparsely spread out across the country.

This thesis will go through the background information, motivation, and related work in Section 1. It will then discuss the theory and details of the correlation techniques used in Section 2. Section 3 will begin to show the application and results of the aforementioned methods. It will also show how we further utilize our results to determine the appropriate size of the training set necessary to avoid any inherent bias in the images. We will conclude in Section 4 and propose future work on this problem in Section 5.

1.1 Related Work

The work presented in this paper is primarily related to two areas of research in computer vision; here we will present some of the work related to algorithms designed to operate on webcam image sequences as well as the use of CCA and other correlation techniques to extract external signals from time-varying image sequences.

TODO

1.2 Background Information

The Archive of Many Outdoor Scenes (AMOS) database [3] has been collecting images from 835 webcams every 30 minutes since March 2006 and now contains over 40 million images. The AMOS dataset is unique in providing time-stamped images from many cameras around the world. No other dataset provides the broad range of geographic locations and the long temporal duration.This database is the largest known collection of natural scenes collected from static cameras and as such offers a wealth of data to test our methods against. While there are cameras located across the world, we focus on those located within the continental United States so that we can collect accurate ground truth weather data.

Our weather data is collected from the Historical Weather Data Archives (HWDA) [8] which is maintained by the National Oceanic and Atmospheric Administration (NOAA), which is an official government organization. The archives maintain a large variety of weather data from the since January 1, 1933 through present day on just over 6,000 weather stations across the continental United States. The data collected includes, but is not limited to, wind speed & direction, precipitation, temperature, relative humidity, vapor pressure, cloud conditions, dew point, and various aggregated data signals.

Chapter 2

Canonical Correlation Analysis (CCA)

The main correlation technique which we will explore is a method called Canonical Correlation Analysis (CCA) [2]. The goal of CCA is to find two transformation matrices A and B to maximize the correlation between two independent data sets $X \in \mathbf{R}^{x \times n}$ and $Y \in \mathbf{R}^{y \times n}$. In other words, CCA looks to find A and B such that $XA \approx YB$. CCA is a way of finding a linear relationship between two independent, multidimensional variables.

MORE

What makes CCA different from other correlation methods is its invariance to affine transformations of the input variables. In other words, CCA will be able to find a linear relationship between two multidimensional variables even if different coordinate systems are used for each variable [1]. This makes CCA a very desirable method for relating image data to environmental signals, which are clearly not measured in the same coordinate system. One very crucial component of CCA is that while the input matrices X and Y can have a varying number of dimensions, they must have *exactly* the same number of samples in order to find an accurate correlation between the two time-varying signals.

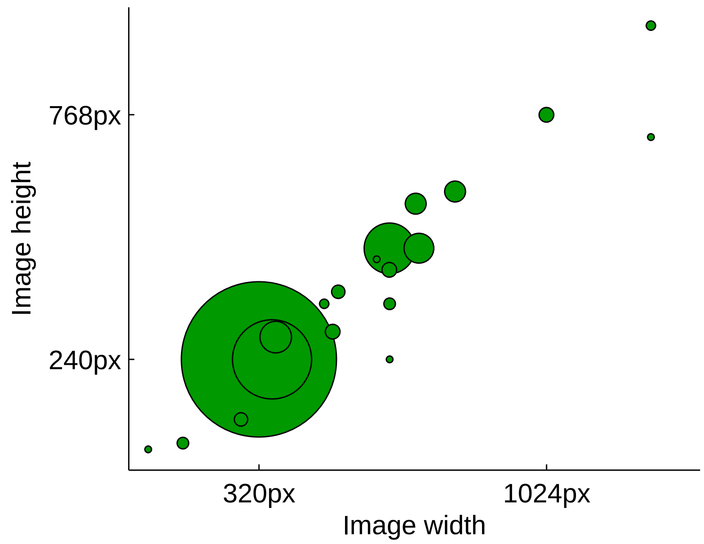

Figure 2.1: The distribution of image sizes, measured in pixels. The circles are centered at the width and height of the image and their sizes are proportional to the number of webcams which output images of that size in the AMOS dataset.

2.1 Applications

We will now focus on the application of CCA to predicting time-varying weather signals from image sequences over the same time period. Given the localized nature of weather data, we assume that the weather station used for ground truth weather data is located near to the camera in order to maximize the accuracy of our predictions. The algorithm takes as input a set of images $I = i_1 \dots i_a$ and a set of weather observations $W = w_1 \dots w_b$. The method will assume the availability of images and weather data with corresponding timestamps. In order to ensure that this invariant holds, the first step of our algorithm is to run through both data sets and remove entries that do not have a corresponding entry in the other dataset. This step guarantees that all of our data samples match and that there are an equal number in both sets, which is required for CCA. Our input is now of the form $I = i_1 \dots i_n$ and $W = w_1 \dots w_n$.

Once our datasets are properly aligned, we turn our attention to the images. Figure 2.1 shows us that the most common size image from a webcam in the AMOS dataset is 320×240 pixels, which means that we can express n image as a $n \times 76800$ matrix. This is clearly a very costly and inefficient way to store image data, especially when we are looking to run CCA with a few hundred images. In order to reduce the storage size for each image, and as a result accelerate the runtime of our algorithms, we will run Principal Component Analysis (PCA) on the images as a way to extract the k most important features from the images and then express each image as a linear combination of these features (we will use k=10). PCA will extract significant scene variations from the set of input images which maximize the covariance of the dataset. We will then look to use these coefficients as input to CCA. PCA will take as input a set of images I and will return three matrices $U \in \mathbf{R}^{m\times k}$, $S \in \mathbf{R}^{k\times k}$, and $V \in \mathbf{R}^{k\times n}$ where m is the length of a single image when expressed as a vector. U contains the k feature vectors, S is a diagonal matrix, and V contains the coefficients of each of the k feature vectors for each of the n images. We can thus create reconstructions of the images by multiplying U, S, and V back together.

The PCA coefficients for each image stored in V and our corresponding weather data $W \in \mathbf{R}^{y\times n}$ will be our input matrices for CCA.

Chapter 3

Results & Analysis

3.1 Wind Speed & Direction

3.2 Vapor Pressure

3.3 Error Analysis

Chapter 4

Conclusion

Chapter 5

Future Work

Appendix A

Camera Information

Appendix B

Weather Data

References

[1] M. Borga, H. Knutsson, and T. Landelius. Learning canonical correlations. In *SCIA10*, Lappeenranta, Finland, June 1997. SCIA.

[2] H. Hotelling. Relations between two sets of variates. In *Biometrika*, volume 28, pages 321–377, 1936.

[3] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Consistent temporal variations in many outdoor scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, MN, June 2007.

[4] Nathan Jacobs, Nathaniel Roman, and Robert Pless. Toward fully automatic geo-location and geo-orientation of static outdoor cameras. In *IEEE Workshop on Applications of Computer Vision (WACV 2008)*, Copper Mountain, CO, January 2008.

[5] Nathan Jacobs, Scott Satkin, Nathaniel Roman, Richard Speyer, and Robert Pless. Geolocating static cameras. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, Rio De Janiero, Brazil, October 2007.

[6] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. Multimedia content processing through cross-modal association. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 604–611, New York, NY, USA, 2003. ACM.

[7] S.G. Narasimhan, C. Wang, and S.K. Nayar. All the Images of an Outdoor Scene. In *Europian Conference on Computer Vision (ECCV)*, volume III, pages 148–162, May 2002.

[8] National Oceanic and Atmospheric Administration. Historical weather data archives, April 2009. http://data.nssl.noaa.gov/.

Vita

Richard Speyer

Date of Birth	October 11, 1986
Place of Birth	Johannesburg, South Africa
Degrees	B.S. Computer Science, May 2009 M.S. Computer Science, May 2009
Professional Societies	Association for Computing Machinery Sigma Xi Scientific Research Society
Publications	N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *Proc. IEEE International Conference on Computer Vision*, Oct. 2007.

May 2009

Predicting Weather from Image Sequences, Speyer, M.S. 2009