

Dimensionality Reduction

Dr Muhammad Atif Tahir
NUCES - FAST

Review Lecture 8

- Association Rules Mining
- Apriori Principle
- FP Tree

Dimensionality Reduction

Dimensionality Reduction

- The input space of many learning problems is of high dimensionality
- This has computational implications and, it makes finding the intrinsic information content difficult
- Dimensionality reduction methods usually try to address these two problems at the same time
- Context:
 - unsupervised learning
 - given a collection of data points in an n -dimensional space
 - a **good representation** of the data in r -dimensional

Dimensionality Reduction - Principle

- **m** examples; **n** dimensional space

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1m} \\ \mathbf{x}_{21}, \mathbf{x}_{22}, ..., \mathbf{x}_{2m} \\ \\ \mathbf{x}_{n1}, \mathbf{x}_{n2}, ..., \mathbf{x}_{nm} \end{bmatrix}$$

Columns are the examples.

This can always be presented by factoring the data matrix as a product of r basis vectors and m code words

When $r=n$, the code words are identical to the original data points

Dimensionality Reduction - Principle

- **m** examples; **n** dimensional space

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)^T$$

[illegible]

Columns are the examples.

This can always be presented by factoring

the data matrix as a product of **r** basis vectors and **m** code words

When $r=n$, the code words are identical to the original data points

Data Encoding

- **m** examples; **n** dimensional space

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$$

$$\begin{array}{ccccc} \left[\begin{matrix} \mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1m} \\ \mathbf{x}_{21}, \mathbf{x}_{12}, ..., \mathbf{x}_{1m} \\ \\ \mathbf{x}_{n1}, \mathbf{x}_{n2}, ..., \mathbf{x}_{nm} \end{matrix} \right] & = & \left[\begin{matrix} ... \mathbf{w}_1^T ... \\ ... \mathbf{w}_2^T ... \\ \\ ... \mathbf{w}_r^T ... \end{matrix} \right] & \mathbf{x} & \left[\begin{matrix} \mathbf{v}_{11}, \mathbf{v}_{12}, ..., \mathbf{v}_{1m} \\ \mathbf{v}_{21}, \mathbf{v}_{12}, ..., \mathbf{v}_{1m} \\ \\ \mathbf{v}_{r1}, \mathbf{v}_{r2}, ..., \mathbf{v}_{rm} \end{matrix} \right] \\ \mathbf{n \times m} & & \mathbf{n \times r} & & \mathbf{r \times m} \end{array}$$

$$\min_{\mathbf{V}, \mathbf{W}} \|\mathbf{X} - \mathbf{WV}\|^2$$

Data Encoding

$$\min_{\mathbf{V}, \mathbf{W}} \|\mathbf{X} - \mathbf{WV}\|^2$$

- Given a new basis \mathbf{W} , each data point is represented as a linear combination of the basis vectors and the goal is to minimize the reconstruction error
- Methods differ in the constraints they impose on \mathbf{V} (the encoding matrix). Consequently, they differ computationally

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \dots \\ \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} \mathbf{W}_{11} \dots \mathbf{W}_{1n} \\ \mathbf{W}_{21} \dots \mathbf{W}_{2n} \\ \dots \\ \mathbf{W}_{n1} \dots \mathbf{W}_{nn} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \\ \dots \\ \mathbf{V}_n \end{pmatrix}$$

Dimensionality Reduction Techniques

Visualize, categorize, or simplify large datasets.

- **Principal Component Analysis**: Finds the dimensions that capture the most variance
- **MDS**: Finds data points in lower dimensional space that best preserves the inter-point distance.
- **Isomap**: Estimates the distance between two points on a manifold by following a chain of points with shorter distances between them. (More accurate in representing global distances than LLE; slower than LLE)
- **LLE** (Local Linear Embedding): Only worries about representing the distances between local points. Faster than Isomap (no worry about global distances).
- **Hessian Eigenmaps**
- **Laplacian Eigenmaps**
- **Charting**
- **SOM** (Kohonen's Self-organizing map)

Feature Selection

Feature Selection

- In many applications, we often encounter a very large number of **potential features** that can be used
- Which **subset of features** should be used for the best classification?
- Need for a small number of discriminative features
 - To avoid “curse of dimensionality”
 - To reduce feature measurement cost
 - To reduce computational burden
- Given an $n \times d$ pattern matrix (n patterns in d -dimensional feature space), generate an $n \times m$ pattern matrix, where $m \ll d$

Feature Selection vs. Extraction

- Both are collectively known as dimensionality reduction
- **Selection**: choose a **best** subset of size m from the available d features
- **Extraction**: given d features (set Y), **extract** m new features (set X) by **linear or non-linear combination** of all the d features
 - Linear feature extraction: $X = TY$, where T is a $m \times d$ matrix
 - Non-linear feature extraction: $X = f(Y)$
- New features by extraction may not have physical interpretation/meaning
- Examples of linear feature extraction
 - Unsupervised: PCA; Supervised: LDA/MDA
- Criteria for selection/extraction: either improve or maintain the classification accuracy, simplify classifier complexity

Feature Selection

- How to find the **best** subset of size m ?
- Recall, **best** means classifier based on these m features has the lowest probability of error of all such classifiers
- Simplest approach is to do an **exhaustive search**; computationally prohibitive
 - For $d=24$ and $m=12$, there are about 2.7 million possible feature subsets! Cover & Van Campenhout (IEEE SMC, 1977) showed that to guarantee the best subset of size m from the available set of size d , one must examine all possible subsets of size m
- Heuristics have been used to avoid exhaustive search
- How to evaluate the subsets?
 - Error rate; but then **which classifier** should be used?
 - Distance measure; Mahalanobis, divergence,...
- Feature selection is an optimization problem

Feature Selection: Evaluation, Application, and Small Sample Performance (Jain & Zongker, IEEE Trans. PAMI, Feb 1997)

- Value of feature selection in combining features from different data models
- Potential difficulties feature selection faces in small sample size situation
- Let Y be the original set of features and X is the selected subset
- Feature selection criterion function for the set X is $J(X)$; large values of J indicates better feature subset; problem is to find subset X such that

$$J(X) = \max_{Z \subseteq Y, |Z|=d} J(Z)$$

Taxonomy of Feature Selection Algorithms

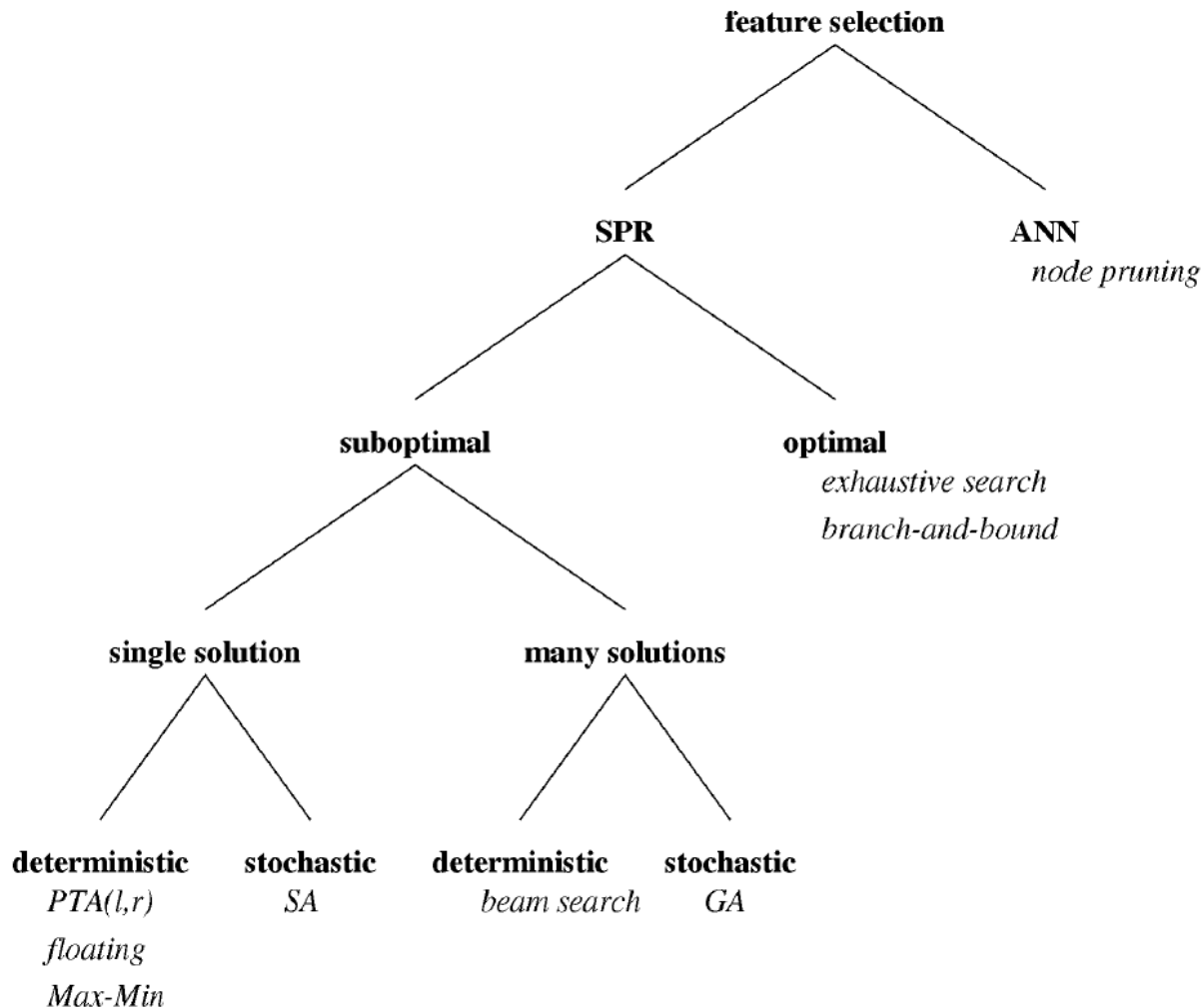


Fig. 1. A taxonomy of feature selection algorithms.

Deterministic Single-Solution Methods

- Begin with a single solution (feature subset) & iteratively add or remove features until some termination criterion is met
- Also known as **sequential methods; most popular**
 - **Bottom up/forward methods:** begin with an empty set & add features
 - **Top-down/backward methods:** begin with a full set & delete features
- Since they do not examine all possible subsets, no guarantee of finding the optimal subset
- Pudil introduced two floating selection methods: **SFFS, SFBS**
- 15 feature selection methods listed in Table 1 were evaluated

TABLE 1
FEATURE SELECTION ALGORITHMS USED
IN EXPERIMENTAL EVALUATION

SFS	SBS	GSFS(2)	GSBS(2)
GSFS(3)	GSBS(3)	SFFS	SFBS
PTA((1), (2))	PTA((1), (3))	PTA((2), (3))	
BB	MM	GA	NP

Sequential Forward Selection (SFS)

- Start with empty set, $X=0$
- Repeatedly add most significant feature with respect to X
- *Disadvantage: Once a feature is retained, it cannot be discarded; nesting problem*

Sequential Backward Selection (SBS)

- Start with full set, $X=Y$
- Repeatedly delete least significant feature in X
- *Disadvantage: SBS requires more computation than SFS; Nesting problem*

Multiclass, Structured and Multi-label Classification

Basic Classification in ML

Input

$\mathbf{x} \in \mathcal{X}$

Output

$y \in \mathcal{Y}$

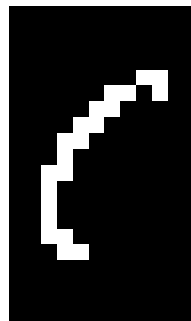
Spam
filtering



Binary



Character
recognition



Multi-Class

C

Structured Classification

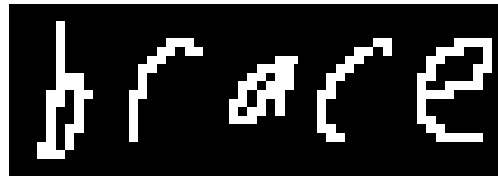
Input

$$\mathbf{x} \in \mathcal{X}$$

Output

$$\mathbf{y} \in \mathcal{Y}$$

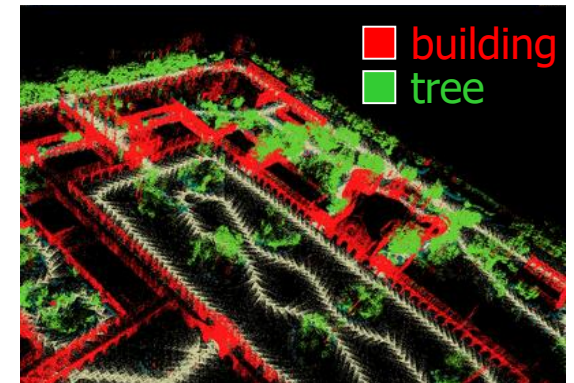
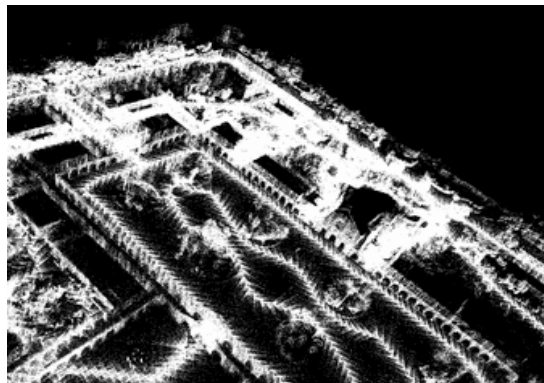
Handwriting
recognition



Structured output

brace

3D object
recognition



Multi-label Classification

Input

$\mathbf{x} \in \mathcal{X}$

Output

$\mathbf{y} \in \mathcal{Y}$



Horse, Person, Tree

Text Mining

(From Book, Chapter 15)


Image / Video Retrieval

Introduction

PONI Advanced Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

https://poni.smu.edu/cgi-bin/Pwebrecon.cgi?PAGE=bbSearch&SE Google

 **PONI** Public Online Information
The Southern Methodist University Libraries Catalog

[Start Over](#)
[Contact the Libraries](#)

[New Search](#) [Headings](#) [Titles](#) [Search History](#) [PONI Databases](#) [Online Resources](#) [Local Libraries](#) [Interlibrary Loan](#) [My Library Account](#) [Request](#) [Help](#)

Database Name: Southern Methodist University Libraries

Basic **Advanced** **Course Reserves**

Search for: video all of these Search in: Keyword Anywhere

☒ AND ☐ OR ☐ NOT

Search for: sound all of these Search in: Keyword Anywhere


☒ AND ☐ OR ☐ NOT

Search for: clip all of these Search in: Keyword Anywhere

50 records per page Search Reset Select More Search Limits

[New Search](#) [Headings](#) [Titles](#) [Search History](#) [PONI Databases](#) [Online Resources](#) [Local Libraries](#) [Interlibrary Loan](#) [My Library Account](#) [Request](#) [Help](#)

[Contact the Southern Methodist University Libraries](#)


top

Done poni.smu.edu

Introduction (cont.)

- Other search engines still using description search method.
- Current image search method: by description.

Introduction (cont.)

Google Advanced Video Search - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://video.google.com/videoadvancedsearch

Google


Sign in

Google
Video BETA

Advanced Video Search

Find results	with all of the words	<input type="text"/>	Google Search
	with the exact phrase	<input type="text"/>	
	with at least one of the words	<input type="text"/>	
	without the words	<input type="text"/>	
Language	Return videos in	Any Language	
Duration	Return videos within the duration	All durations	
Price	Return videos with the price of	All	
Domain	<input type="button" value="Only"/> return videos from the site or domain	<input type="text"/>	
		e.g. youtube.com	
Genre	Return videos from	<input checked="" type="radio"/> all genres <input type="radio"/> specific genres	

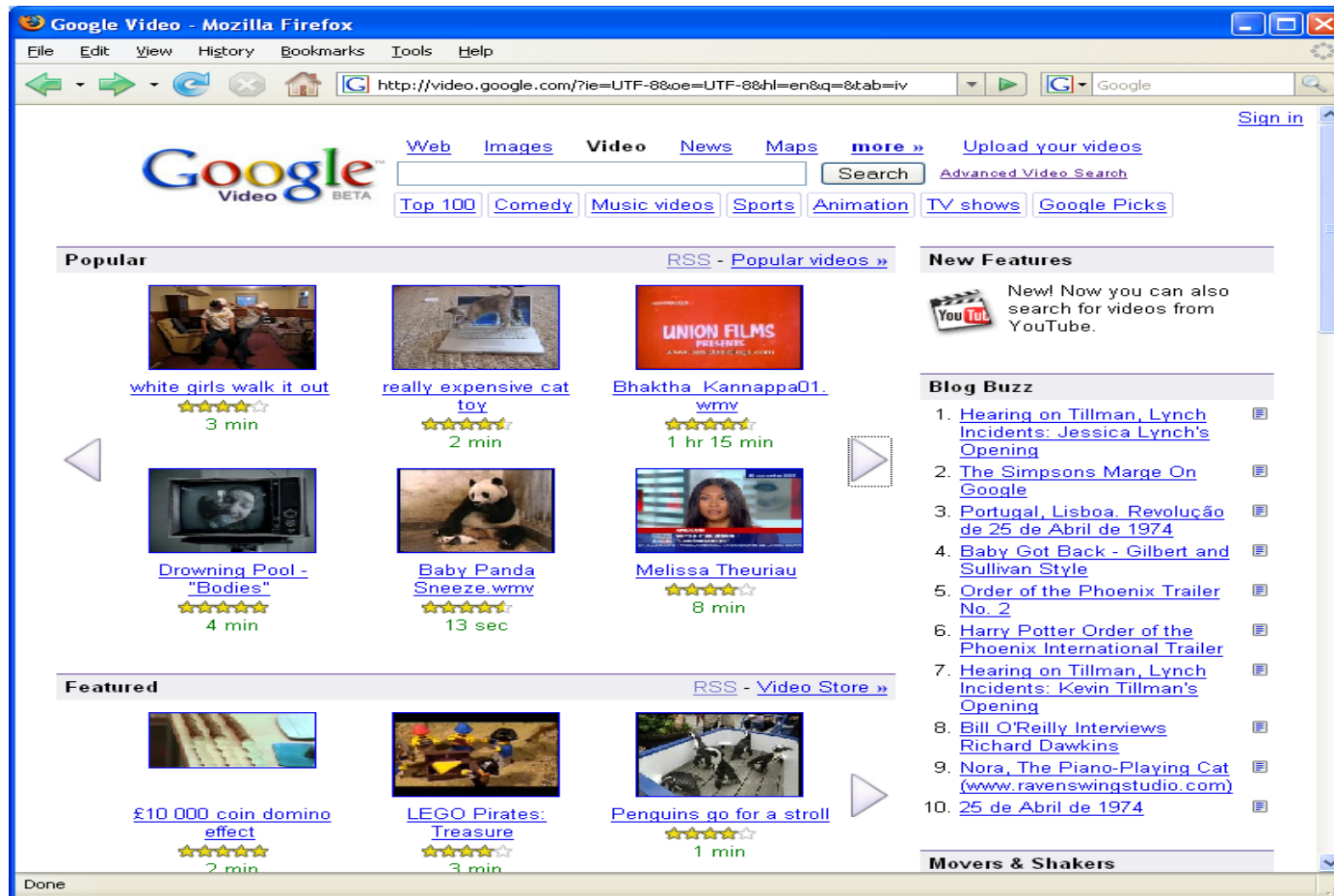
Sort results by	Relevance
Results per page	10 results

Also try our international versions:
[U.S.](#) - [Australia](#) - [Canada](#) - [Deutschland](#) - [España](#) - [France](#) - [Italia](#) - [Nederland](#) - [Polska](#) - [U.K.](#)
[About RSS](#)  - [Discuss](#) - [Download player](#) - [Terms](#) - [Help](#) - [About Google Video](#)

©2007 Google

Done

Introduction (cont.)



Introduction (cont.)

- Picture is worth a thousand words.
- More than words can express.
- Growing number video clips on MySpace and YouTube, there is a need for a video search engine.

Introduction (cont.)

YouTube - Broadcast Yourself. - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.youtube.com/browse?s=mp

Sign Up | My Account | History | Help | Log In

Search

Videos Categories Channels Community Upload Videos

Videos

Most Viewed (Today)

My Videos - Favorites - Playlists - Inbox - Subscriptions

Pages: 1 2 3 4 5 Next

Ads by Goooooogle

Browse

- [Most Recent](#)
- Most Viewed**
- [Top Rated](#)
- [Most Discussed](#)
- [Top Favorites](#)
- [Most Linked](#)
- [Recently Featured](#)
- [Most Responded](#)
- [Watch on Mobile](#)

Time

- Today**
- [This Week](#)
- [This Month](#)
- [All Time](#)

Category

- All**
- [Autos & Vehicles](#)
- [Comedy](#)
- [Entertainment](#)
- [Film & Animation](#)
- [Gadgets & Games](#)
- [Howto & DIY](#)
- [Music](#)
- [News & Politics](#)
- [People & Blogs](#)
- [Pets & Animals](#)
- [Sports](#)
- [Travel & Places](#)

Language

- English**
- [Spanish](#)
- [Japanese](#)
- [German](#)
- [Chinese](#)

The Most Amazing Basketball Shot EVER
00:17
Added: 1 day ago
From: [kasnewalo](#)
Views: 259,750
★★★★★
783 ratings

15 year old tries to out drive the police at 150MPH!
04:45
Added: 1 day ago
From: [kasnewalo](#)
Views: 189,132
★★★★★
765 ratings

Don't Play Entropia Universe
00:17
Added: 12 hours ago
From: [EUWarning](#)
Views: 153,645
★★★★★
100 ratings

LINK LIVEFOOTSTREAM S CHARGES £5 FOR - FREEHERE NO CHARGE!
03:12
Added: 1 day ago
From: [AndAgain432](#)
Views: 117,313
★★★★★
116 ratings

Prepare to be Shocked
You may be younger than you think. Take the RealAge test and find out.
[www.RealAge.com](#)

Weight Loss For Men
The top 10 things guys need to know about fitness. Free tips at AskMen.
[www.AskMen.com](#)

Exhausted All The Time?
It's Not Your Fault. You Just Need To Boost Your HGH Levels. It's Easy
[www.HGH-Facts.com](#)

Copy your iPod music
Transfer songs from your iPod or iPod Mini to your computer.
[www.findleydesigns.com](#)

ZUNE Unlimited Downloads
Music, Movies, MP4 Videos, Sports Instant Access for Only \$38.96
[www.ZuneReactor.com](#)

4 year old child injured at Colorado state football practice
00:34
Added: 1 day ago
From: [kimmelscorner](#)
Views: 108,910
★★★★★
161 ratings

Order of the Phoenix Trailer No. 2
02:12
Added: 1 day ago
From: [leakynewsdotcom](#)
Views: 102,481
★★★★★
395 ratings

Top 10 Tips to Get Bathing Suit Ready: sparkpeople.com top10
02:11
Added: 1 day ago
From: [elfersp](#)
Views: 100,358
★★★★★
88 ratings

Harry Potter and the Order of the Phoenix Domestic Trailer
02:13
Added: 1 day ago
From: [mvideoes](#)
Views: 99,230
★★★★★
179 ratings

LIVEFOOTSTREAM

Scarlett Wines Out

Will Ferrell's landlord

Diet.com Weight

Done

Introduction (cont.)

- Therefore, we need a better search technique – Content-Based Video Retrieval System (CBVR).

Introduction (cont.)

- What good is video retrieval?
 - Historical Achieve
 - Forensic documents
 - Fingerprint & DNA matching
 - Security usage

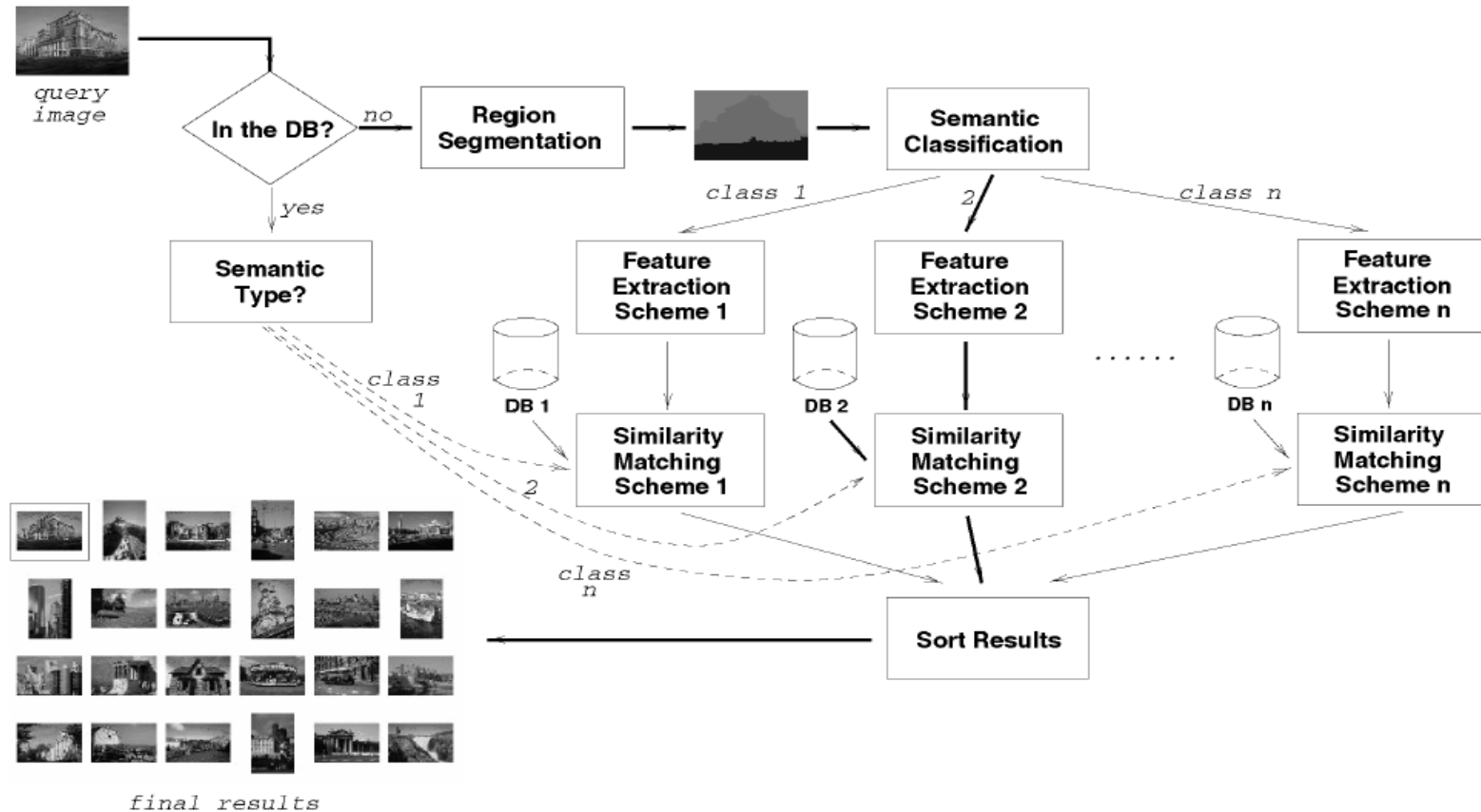
Overview (cont.)

- CBVR has two Approaches:
 - Attribute based
 - Object based
- CBVR can be done by:
 - Color
 - Texture
 - Shape
 - Spatial relationship
 - Semantic primitives
 - Browsing
 - Objective Attribute
 - Subjective Attribute
 - Motion
 - Text & domain concepts

Overview (cont.)

- CBVR has two phases:
 - Database Population phase
 - Video shot boundary detection
 - Key Frames selection
 - Feature extraction
 - Video Retrieval phase
 - Similarity measure

Overview (cont.)



[Wang, Li, Wiederhold, 2001]

References

- Professor Dan Roth Lecture Notes on Dimensionality Reduction, University of Illinois at Urbana Champaign
- Content-Based Video Retrieval System, Lecture Notes by Edmund Liang
- Lecture Notes By Anil Jain (MSU) on Feature Selection

Questions!