ICME 2003

July 8th, 2003

SUPERVISED CLASSIFICATION FOR VIDEO SHOT SEGMENTATION

Yanjun Qi, Alex Hauptmann, Ting Liu
qyj@cs.cmu.edu
Carnegie Mellon University

Roadmap

- Introduction
- Previous work
- Video shot segmentation based on supervised classification
- Experiments
- Summary

Introduction

- Temporal video shot segmentation
  - The first step for automatic video browsing and retrieval
  - Has been extensively studied

- Shot
  - An unbroken sequence of frames taken from one camera
- Shot transitions
  - Two basic types
    - Cut transitions
    - Gradual transitions
  - Gradual transitions are more difficult to detect than cuts

Previous Work

- Most existing algorithms
  - Thresholding differences between successive frames
  - Difficult to get suitable thresholding - sensible to video type

- Among machine learning methods that have been tried
  - K-means to cluster frame differences
  - HMMs with separate states to model shot cuts, fades, dissolves, pans and zooms
  - “Dissolve synthesizer” to create artificial training data for supervised learning methods
  - Statistical detector based on minimization of the average detection-error probability for cuts and dissolves

Video Shot Segmentation Based on Supervised Classification

- Treat video shot segmentation as a categorization task
  - Classify every frame in the video stream into
    - “common shot frame”
    - “cut frame”
    - “dissolve frame”
    - Other transition types such as “fade”, “wipe”, etc.

- Classification framework
  - Use different kinds of video features in an integrated structure
  - Supervised learning enables reliable estimation of thresholds
    - Requires representative training data

System Overview

System processing steps:

(1) Treat every frame as a single feature vector

(2) Classify each frame into exactly one class

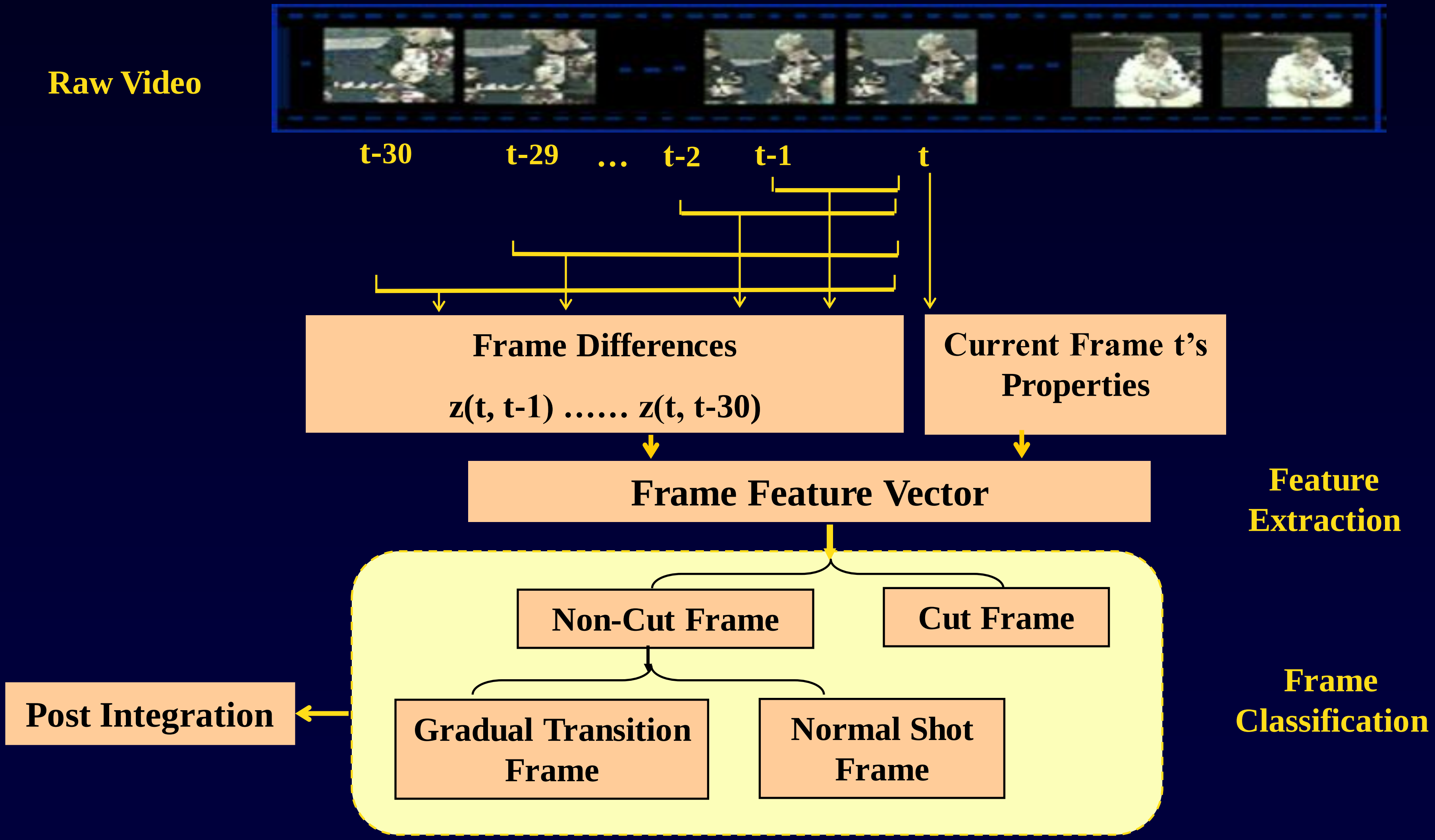
(3) Post processing for the final segmentation result

Our implementation

- Two broad boundary types: *Cuts* and *Gradual Transitions*
  - Capable of detecting many other types of transitions

- Learn to categorize each frame into one of three classes
  - "hard cuts frame"
  - "gradual transitions frame"
  - "common shot frame" (non-boundary)

System Overview (Continued)



Raw Video

t-30 t-29 … t-2 t-1 t

Frame Differences

z(t, t-1) …… z(t, t-30)

Current Frame t's Properties

Frame Feature Vector

Feature Extraction

Non-Cut Frame

Cut Frame

Gradual Transition Frame

Normal Shot Frame

Frame Classification

Post Integration

Step One: Frame Feature Extraction

Find features to reliably distinguish different segmentation classes

- Frame features derived in two ways:
    - Current frame property
    - Frame difference to previous frames

Frame Feature Extraction (Continued)

- Frame Difference
  - Compute differences in a window of 30 frames
    - between frame t and frame t-1, up to frame t and frame t-30
  - Compute a total of 60 differences all in the YUV color space:
    - 30 differences based on Whole-frame color histogram
    - 30 differences from 8*8 block-wise histogram difference of frames

These 60 window-based differences represent a frame's temporal relationship within its neighborhood

- Current Frame Property
  - Camera motion probability
  - Black frame likelihood

Step Two: Frame Classification

- Two – Level Binary Classification
  - First Level: Cut vs Non-Cut
    - Binary classifier to categorize each frame into “non-cut frame” or “cut frame”
  - Second Level: Shot vs Gradual
    - Binary classifier to distinguish a “shot frame” from a “gradual transition frame”

In general, distinguishing cuts from gradual transitions or normal shots is much easier than separating gradual frames from normal shot frames

Frame Classification (Continued)

- Explored three supervised classification methods
  - K-nearest Neighbor (KNN)
    - Classify test vector based on it k nearest neighbors in the training set
  - Naive Bayes Classifier (BC)
    - Use features' joint probabilities to estimate the probabilities of a category given a data point
  - Support Vector Machine (SVM)
    - Based on the structural risk minimization principle
    - Aims to find a decision surface that "best" separates the data points in two classes

Step Three: Post Processing

- Wavelet Smoothing
  - Smooth each non-cut frame‘s classification score
  - Suppress the noise and consolidate the classification scores corresponding to a sequence of gradual transition frames
- Temporal Integration for Gradual Transition
  - Multiple transitions are unlikely to be immediately adjacent to each other

Experiments

Data Corpus

- NIST TREC-2001 Video Track Collection
  - Provides a standard data corpus and unified evaluation criteria
  - Allows consistent and objective comparison of different systems

- Our experiments used 4 hours of video from this corpus, or 13 MPEG-1 video files at slightly over 2GB of data
  - 420,976 frames and 2462 transitions
    - 1670 cuts (67% of all transitions)
    - 792 gradual transitions

Evaluation

- Shot segmentation reference data
  - Constructed manually by NIST
  - Evaluation software provided by NIST

- We use Precision / Recall / F1 score to evaluate
  - **Precision**
    - Among the transitions (cut or gradual) detected by the system, how many are true transitions?
  - **Recall**
    - For all possible transitions (cut or gradual), how many were detected by system?
  - $$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Four Interesting Experimental Runs

- **Run 1 (30.bc.bc)**
  - Only block wise histogram difference (30 features)
  - BC for both levels of classification
- **Run 2 (30.knn.knn)**
  - Block wise histogram difference (30 features)
  - kNN for both levels of classification
- **Run 3 (62.knn.knn)**
  - Global and block-wise histogram differences, camera motion likelihood and black-frame likelihood (30+30+2 features)
  - kNN for both levels of classification
- **Run 4 (62.svm.knn)**
  - Use the same 62 features as Run 3
  - Uses a linear SVM for the first level classification
  - kNN for the second level

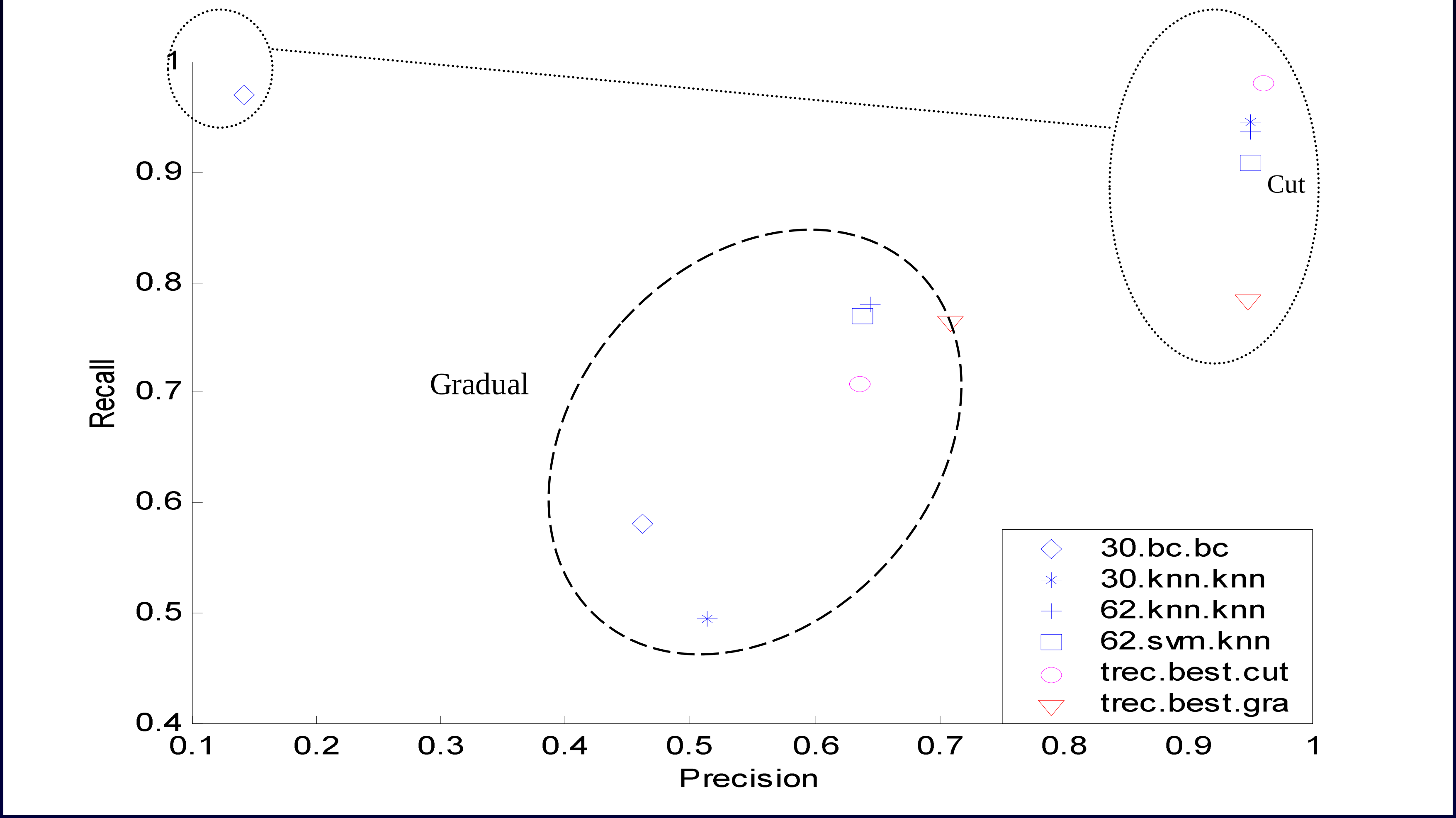
Comparison Results

F1 comparison for these six runs

Runs	cut_f1	gradual_f1	sum_f1
30.bc.bc	0.241644	0.500100	0.3967
30.knn.knn	0.947389	0.485034	0.6700
62.knn.knn	0.942435	0.698285	**0.7935**
62.svm.knn	0.928222	0.685770	0.7828
TrecBestCut (non CMU)	**0.965900**	0.670600	0.7887
TrecBestGra (non CMU)	0.857200	**0.729700**	0.7807

Compared to the best performing systems of 2001 TREC evaluation, our performance was best overall in terms of F1.

Comparison Results (Continued)



Precision vs. Recall for Cuts and Gradual Transitions

Summary

- Transform video shot segmentation to categorization task
  - Unified framework enables use of different types of features
- Supervised classification
  - More reliable estimation than previous threshold-based methods
- Excellent performance on
  - Unified benchmark evaluation
  - Standard TREC 2001 Data Corpus
- The general window-based classification framework could easily be extended to other video analysis tasks

Thank you for your attention

Questions ?