MMAI Term Project

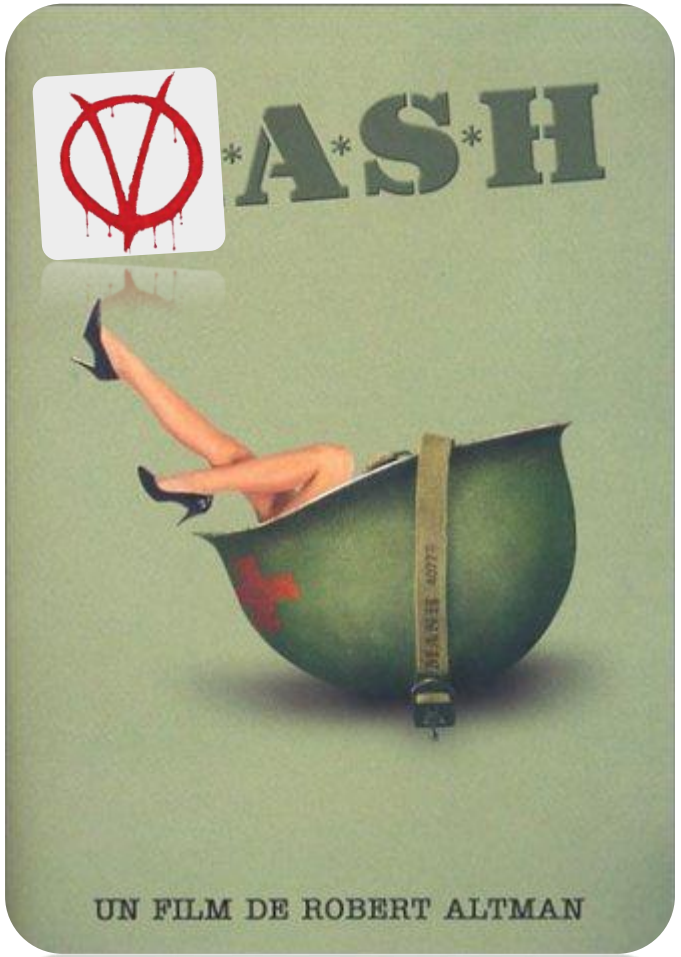
VASH

郝瑞尼　　蔡格昇　　蔡宗諭

The Vision

"Perceptual Movie Identification"

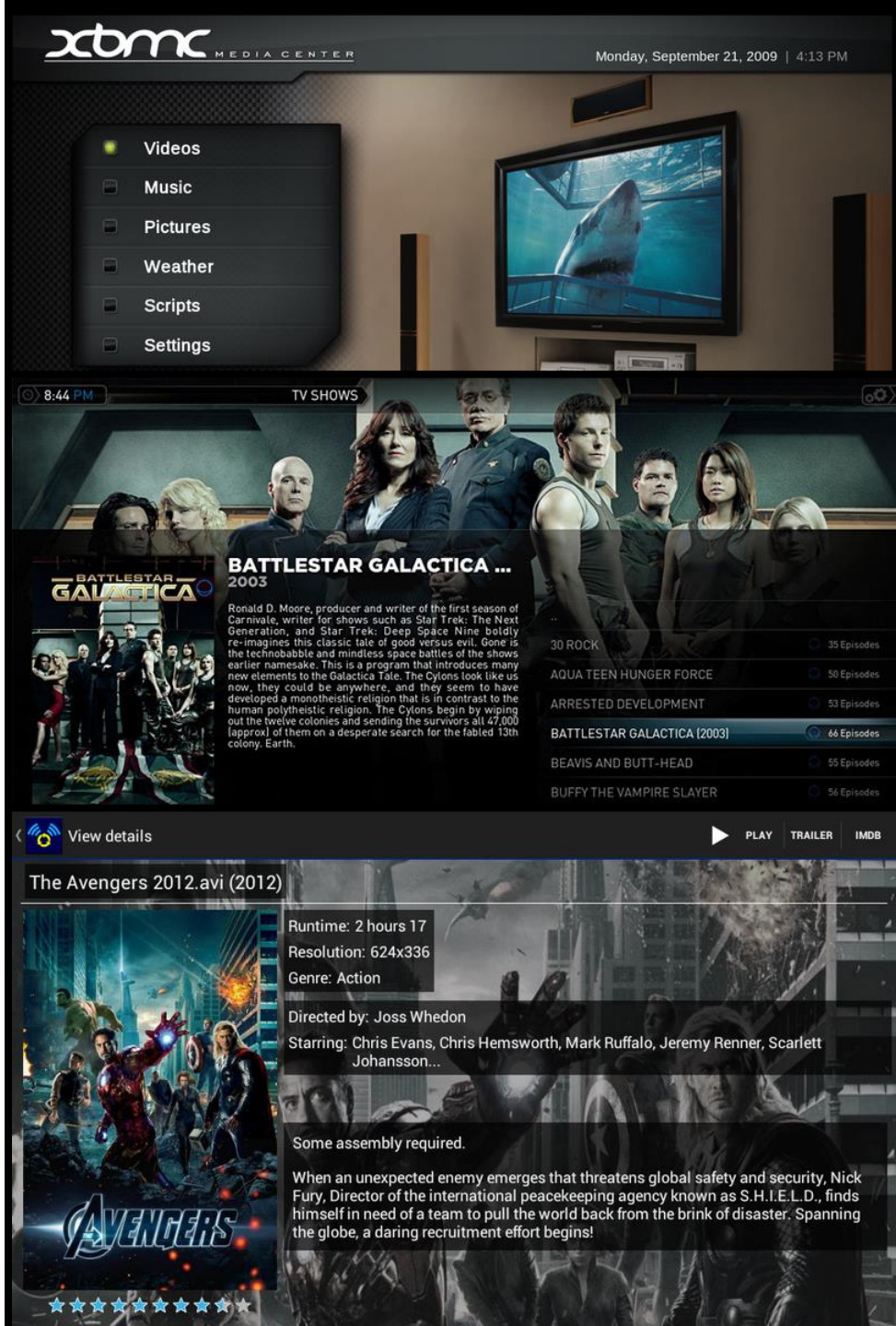
- Look at Movie content
- Identify Movie (name, year, id)
- 100% accuracy and fully automatic



The Why?

- Virtually NO metadata / tagging
- HTPC software used by millions
- Manually organizing media files
- Wasting millions of hours

“ We can do better! “



“

I spent 2 days organizing all my movies, series and documentaries. They were already extremely organized, but not good enough for XBMC. So I created an NFO file for every episode. **Was that enough? No.**

”

douglasrac

Junior Member

Posts: 9

Joined: Sep 2013

Reputation: 0

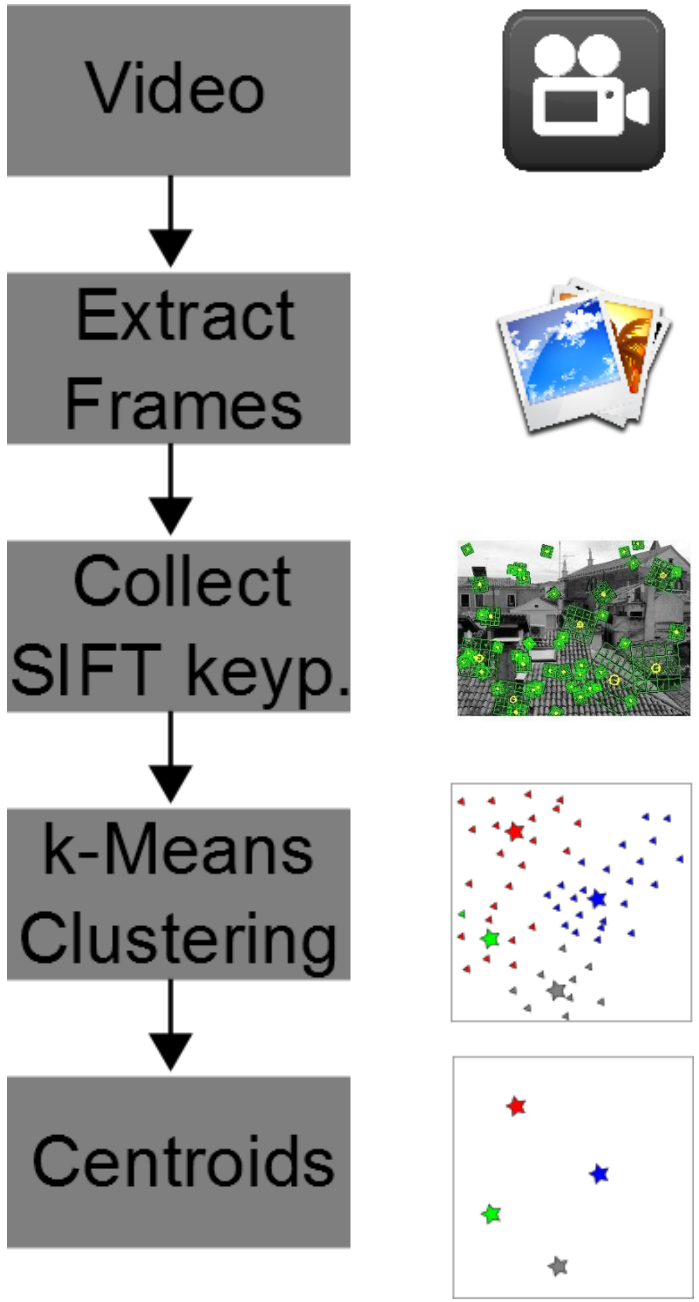
The How

- Sample key frames from video files
- Extract SIFT keypoints for video (sub-)sequences
- Compute signature using Histogram of Visual Words
- Match signature against our reference database
- Return video identity

“ Yang et al., 2007, Evaluating bag-of-visual-words representations in scene classification ”

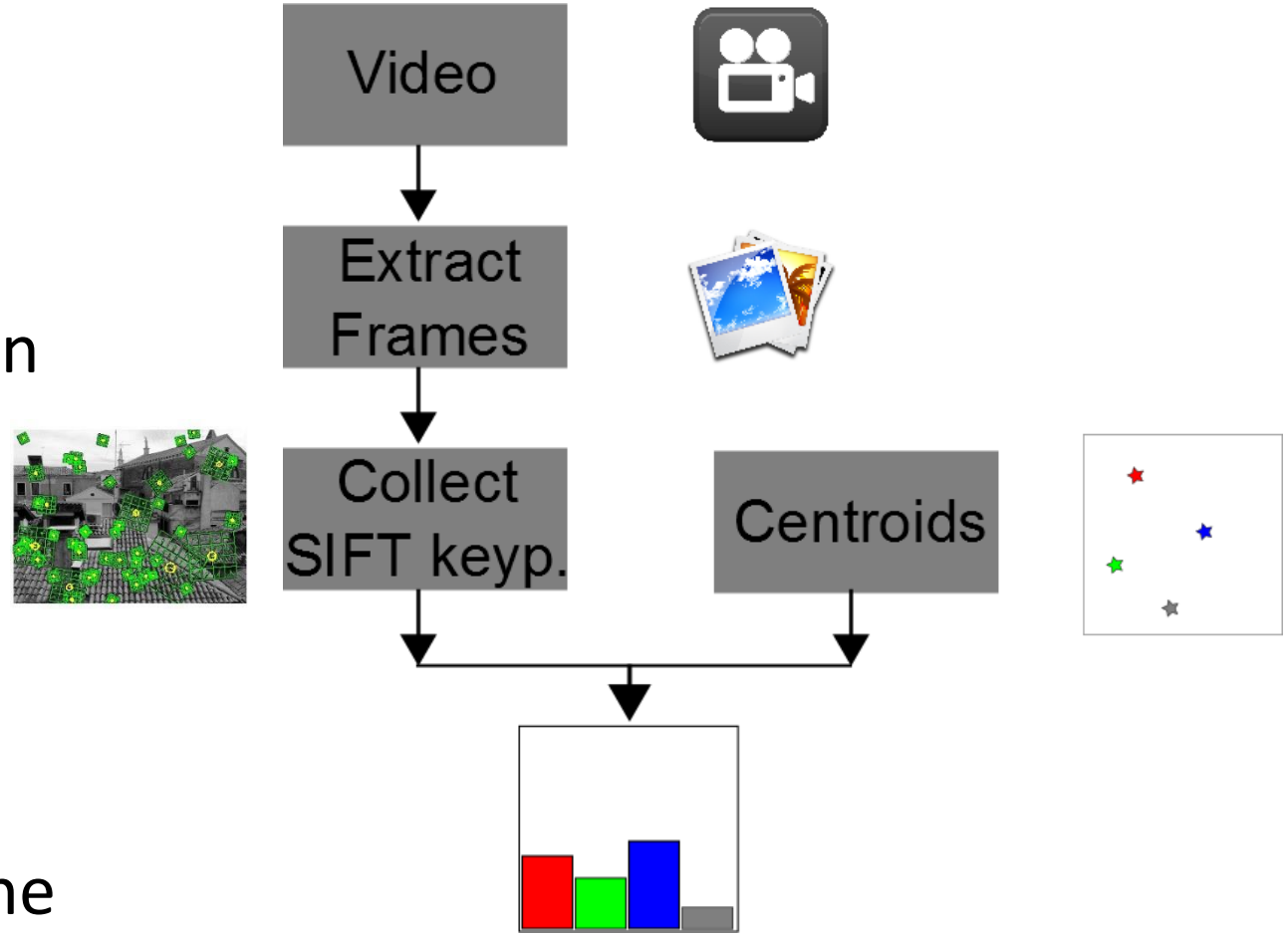
Generation of Visual Word Dictionary

- Select a number of frames from videos
- Compute and aggregate
- SIFT keypoints
- Run k-means clustering
- The resulting k-centroids are the Visual Words



Training Signatures

- Which frames to select?
  - ✓ Choose certain frames at certain intervals
- Compute SIFT key points
- Quantize keypoints to Visual Words
  - ✓ Compute the distances to each centroid and pick the closest one
- Signature is a Histogram of Visual Words, e.g. VW 1 occurred 5 times



How does SIFT work?

- Candidate keypoints are found by searching extrema in the Gaussian-scale space
- Further refinement
- Compute orientation and magnitude
- Combine to keypoint descriptor

“ D. G. Lowe, 2004, Distinctive Image Features from Scale-Invariant Keypoints, Int. Journal of Computer Vision “

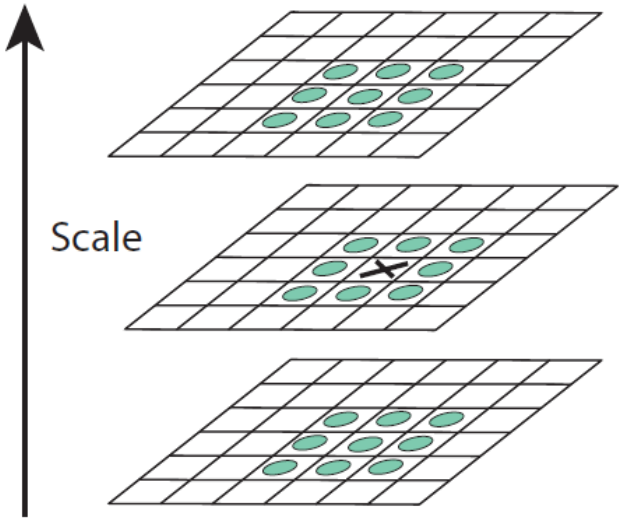
SIFT keypoint discovery

Extrema

- Use a DoG approximation for faster computation

$$\underbrace{G(x, y, k\sigma) - G(x, y, \sigma)}_{\text{DoG}} \approx (k-1)\underbrace{\sigma^2 \nabla^2 G}_{\text{LoG}}$$

- Give the current and two neighboring scales, select a point if all 26 neighbors have smaller or greater value

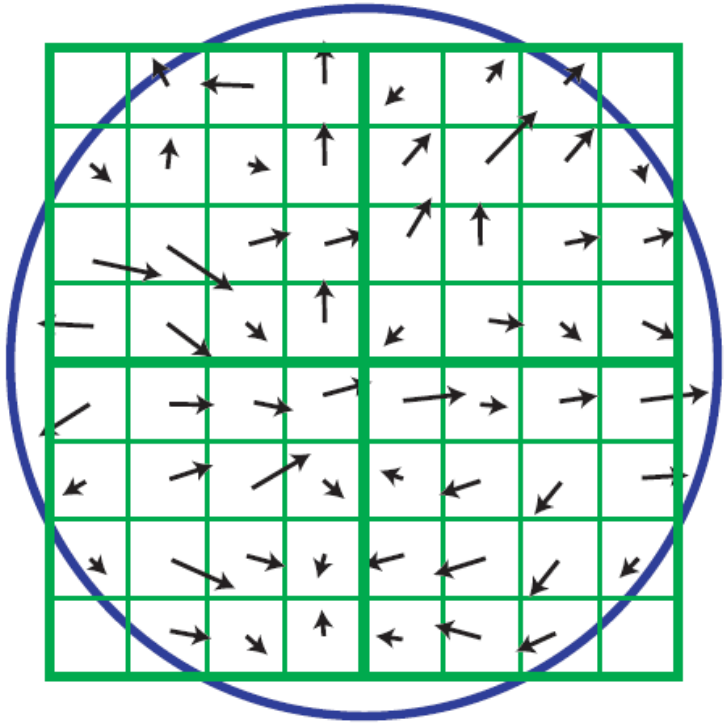


Orientation

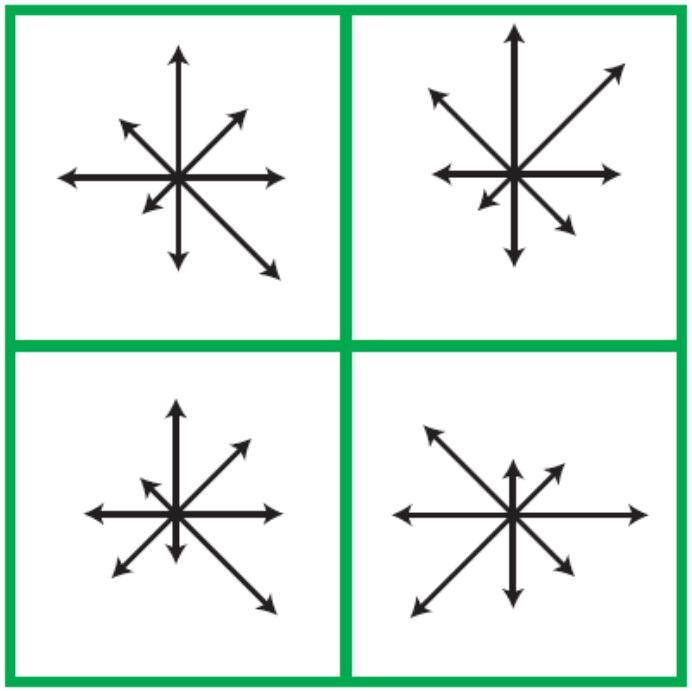
$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2}$$

$$\theta(x, y) = \tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right)$$

SIFT keypoint discovery



The orientations in an area around the keypoint



Binning - Creating a histogram from the orientations in the left figure

Note: In his paper, Lowe recommends 4x4x8 histograms instead of 2x2x8 ones.

Evaluation

DataSet 1: True Blood – Season 5 (12 episodes, real-world data)

DataSet 2: CC_WEB_VIDEO: Near-Duplicate Web Video Dataset

Evaluate with k=100 and k=1000 visual words respectively

Evaluate against 9 attack types (logo, subtitles, crop, etc)

Test each video for each attack (Null-Attack is naturally 100%)

Performance for k=1000 (on Intel Core 2 Duo 2.54 GHz)

- Generate Centroids 43:30 min
- Train Signatures 02:02 min
- Query 9 sec

Test Results

Attack Type	Accuracy k=100	Accuracy k=1000
Heavy Blur	8 %	8 %
Light Blur	38 %	75 %
Motion Blur	8 %	8 %
Radial Blur	8 %	8 %
Crop	100 %	100 %
Logo	100 %	100 %
Heavy Sharpen	33 %	92 %
Light Sharpen	42 %	100 %
Subtitles	96 %	100 %

k … size of visual word vocabulary

Future Works

- More evaluation with real world data required
  - Detection of the same movie with slightly different length
  - Detection of the same movie in different codecs (and key frame intervals)
  - Bigger data sets for testing real world attacks
- Distributed collection of vocabulary and signatures
- Fusion of Features (e.g. SIFT features + Color Histogram)

THE END

謝謝

Questions?

Visit us at http://vash.in