Classification of Video Shots Using 3D Camera Motion

Task Definition

- Goal: Given the video sequence categorize video shots into the following three categories:
  1. Pan/Track: horizontal motion
  2. Tilt/Boom: vertical motion
  3. Zoom: either zoom-in or zoom-out
- Testing Dataset: TRECVID 2005 dataset
  1. Broadcast videos
  2. MPEG-1
  3. 140 videos (30mins to 1 hour each), 200~400 shots per video.

Motion Models (long videos)


Pan (right)


Tilt (up)


Zoom (out)


Track (right)


Boom (up)


Static

Motion Models (short videos)


Pan (right)


Tilt (up)


Zoom (out)

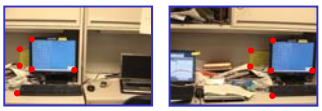

Track (right)


Boom (up)

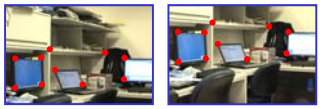

Static

Motion Models

(a). Pan (Left)



(b). Track (Left)



(c). Tilt (Down)

(d). Boom (Down)

(e). Zoom (Out)

(f). Static

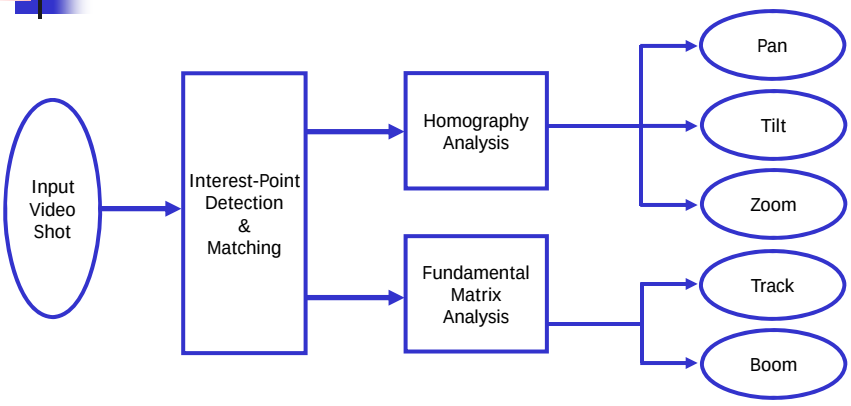
Common Approach

- Based on the analysis of the optical flow field; (2D motion)
- Computationally Expensive;
- Cannot distinguish pan/tilt from track/boom.

Proposed Approach

- Uses 3D geometric analysis
  - Homography: pan, tilt and zoom;
  - Fundamental matrix: track and boom.
- Interest-point correspondence (low computational cost).
- Is able to separate pan from track.

Work Flow



Pin-Hole Camera Model

Projects a 3D point $U=[X,Y,Z,1]$ onto a 2D plane $u=[x,y,1]$

$$u \sim \mathrm{P}U = \mathrm{K}[\mathrm{R}\,|\,\mathrm{v}]U \qquad \mathrm{K} = \begin{bmatrix} f & \gamma & u_0 \\ 0 & \lambda f & v_o \\ 0 & 0 & 1 \end{bmatrix}$$

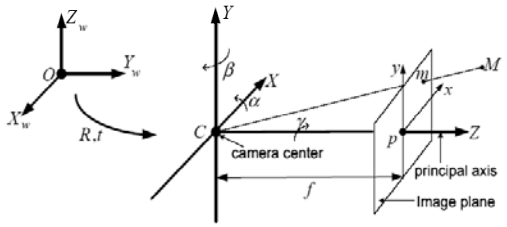
R - 3x3 orthonomal rotation

v - translation vector

f - focal length

λ - aspect ratio

γ - skew factor



Homography Analysis (pan, tilt)

These motions are caused by pure 3D camera rotations.

Before rotation: u After rotation: u'

$u = \mathrm{K}[\mathrm{I}\,|\,0]U$

$u' = \mathrm{K}[\mathrm{R}\,|\,0]U = \mathrm{KRK}^{-1}\mathrm{K}[\mathrm{I}\,|\,0]U = \mathrm{KRK}^{-1}u$

$u' = \mathrm{H}u$ and

H ↔ R (Conjugate with)

same eigenvalues

$\{\mu, \mu e^{i\theta}, \mu e^{-i\theta}\}$

Criteria: eigenvector of the real eigenvalue corresponds to the rotation axis.

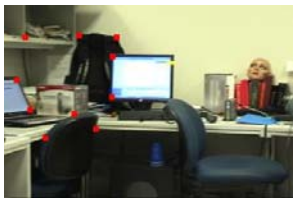
$$v = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

If Pan Motion: $b >> a, c$

If Tilt Motion: $a >> b, c$

Example of Pan Motion

Pan Right



$$H = \begin{pmatrix} 1.1938 & 0.0108 & -235.7323 \\ 0.0289 & 1.1420 & -8.3286 \\ 0.0005 & 0.0000 & 1.0000 \end{pmatrix}$$

Rotation Angle: 16.75°

Rotation Axis:
[0.0007, 1.0000, 0.0000]

Y-Axis

Homography Analysis (zoom)

Before zoom: u After zoom: u'

$$u = \mathrm{K}[\mathrm{I}\,|\,0]\,U$$

$$u' = \mathrm{K}'[\mathrm{I}\,|\,0]U = \mathrm{K}'\,\mathrm{K}^{-1}\mathrm{K}[\mathrm{I}\,|\,0]\,U = \mathrm{K}'\,\mathrm{K}^{-1}u$$

$$u' = \mathrm{H}u$$

$$\mathrm{H} = \mathrm{K}'\,\mathrm{K}^{-1} = \begin{pmatrix} \lambda \mathrm{I} & (1-\lambda)\,p_0 \\ 0^T & 1 \end{pmatrix}$$

p_0 - principal point

$\lambda = f'/f$ - scaling factor

Criteria:

1. Lower triangle of H should be zeros;
2. First two diagonal elements are equal.

Example of Zoom Motion

Zoom In

$$H = \begin{pmatrix} 1.5632 & 0.0036 & -98.3642 \\ 0.0208 & 1.5426 & -64.1500 \\ 0.0002 & 0.0001 & 1.0000 \end{pmatrix}$$

Scale Factor ~= 1.5529

Principal Point (177.9, 116.0)
It's near the image center.
(image size = [360]x[240])

Fundamental Matrix Analysis (track, boom)

These motions are caused by pure 3D camera translations.

- Cannot use this homography:
  1. Homography → scene is not planar.
- When camera moves: using fundamental matrix.

Before motion: $u = \mathrm{K}[\mathrm{I} \mid 0]\, U = \mathrm{P}U$ After motion: $u' = \mathrm{K}[\mathrm{I} \mid \mathrm{v}_0]\, U = \mathrm{P}'U$

$l' = \mathrm{F}u \qquad l = \mathrm{F}^{\mathrm{T}}u'$ where $\mathrm{F} = K'RK^{-1}(e')_{\times}$

l - epipolar line
P^{+} - pseudo-inverse of P
e' - epipole
$(\bullet)_{\times}$ - skew matrix

If only tracking is present: $R = \mathrm{I}$ and $\mathrm{F} = (e')_{\times}$

Criteria: the epipole should be located at the vanishing point of the translation direction.

Example of Track Motion

Track Right

A pair of images from a track-right motion

$$F = \begin{pmatrix} 0.0000 & 0.0012 & -0.2258 \\ -0.0012 & 0.0000 & -0.9742 \\ -0.2258 & -0.9742 & 0.0000 \end{pmatrix}$$

Epipole Location:
(0.9742, -0.2258, -0.0012)

Almost located at the infinity, the vanishing point of the x-axis.

Refinement & Ranking

Major mis-classifications: caused by shaking of the camera.

A true pan/track shot should have:

- A relative velocity in the moving direction;
- Fewer change count in the moving direction change;
- Lower ratio between direction change and overall displacement.

Ranking function:

$$f \propto \max\left(vel^{+}, vel^{-}\right) - \min\left(vel^{+}, vel^{-}\right)$$

$$\propto \left(\frac{C_{change}}{Length}\right)^{-1}$$

$$\propto \left(\frac{Accu.\ Change}{Overall\ \ Displacement}\right)^{-1}$$

Summary of Algorithm

1. Sub-sample the video sequence.
2. Detect the SIFT interest points in each frame of sampled video and determine the correspondences.
3. For each pair of adjacent images, do
   - Compute homography using RANSAC;
   - Compute fundamental matrix using RANSAC;
   - Determine motion type.
4. Compute the percentage of each motion type in the shot.
5. Rank each shot according to each motion category.
6. Return the ranking list of each motion type.

System Performance

1. Proposed method has been tested on the TRECVID 2005 dataset;
   - MPEG-1 Videos;
   - 140 Videos with 200~300 shots each.
2. Results have been ranked based on the proposed functions.
3. Performance:
   - Manually evaluated first 200 returns of each motion: >= 85% precision.
   - Between 200~1500 returns, manually evaluated every 10th return: ~= 80% precision.
4. Results have been submitted to NIST for evaluation.
   - Expected return date is Sept. 20.