Browsing: Query Refinement and Video Synopsis

Yonatan Bisk

April 23, 2009
CuZero: Frontier of Interactive Visual Search

Graph-Cut Transducers for Relevance Feedback

Non-chronological Video Synopsis and Indexing

Conclusion
Goals
Finding interesting content in video and images

Solution
- Better video and image search
- Video synopsis to quickly scan long video
Current Situation

Need to shorten clip without cutting too many frames
Need to only cut out “unimportant” frames
Need to handle lighting and scenery changes
Need to handle never ending video
Current Situation - Video Search

Query: “you’re yes and you’re no you’re up and you’re down”

Did you mean: you are yes and you’re no you’re up and you’re down

KJ-52 & Blanca Reyes - You’re Gonna Make It
03:47 - 1 year ago
And help you and plus give you strength too you're gonna make it man you're gonna be ok dude ... KJ-52 you' YouTube.com

How To Know You’re In Love
01:43 - 7 months ago
You’re willing to put up with her boring family. You love spending time with her, regardless of what you’re doing. And you ... videomug.com

Katy Perry Hot N Cold Sex and the City Diva Australia
15:49 - 1 month ago
then you're cold you're yes then you're no you're in then you're out. You're up and then you' Flirtsz.net

Italian promo Katy Perry - Hot n Cold
15:09 - 25 days ago
then you're cold you're yes then you're no you're in and you're out. You're up and you' V4 drunksexs.net

The Affluents - Stop What You're Doing (PART I)
09:38 - 1 year ago
say you in the car you little soldier you with the scars you got the power you shooting star open your eyes and stop what you're doing you YouTube.com

cold lyrics with n hat
12:51 - 1 month ago
then you're cold you're yes then you're no you're in and you're out. You're up and you'
Minimal Search Requirements

Video
- Need to know content of video/images
- Need to understand dialog (video)
- Need to have results containing all arguments
- Allow user to specify they mean “real” animals
- Specify view of object/animal they are interested in (images)
Current Situation - Image Search

Yonatan Bisk

Browsing: Query Refinement and Video Synopsis
Outline

CuZero: Frontier of Interactive Visual Search

Graph-Cut Transducers for Relevance Feedback

Non-chronological Video Synopsis and Indexing

Conclusion
Employs a unique query process that allows zero-latency query formation for an informed human search. Relevant visual concepts discovered from various strategies are automatically recommended in real time. Also introduces a new intuitive visualization system.
Demo

GeoTag Columbia
CuZero: Frontier of Interactive Visual Search
Graph-Cut Transducers for Relevance Feedback
Non-chronological Video Synopsis and Indexing
Conclusion

Average Precision comparison

No user provided labels and performed in 1/3 the time
Summary

- Combine existing conceptual resources
- Use concept information to assist in query formation
- Visualize results
- Plot results to allow for combining concepts
- Allow for advanced queries to form (geo info, etc.)
Pluses

- Zero latency process to aide in query formation
- Interactively choose best query suggestion
- Demonstrates interactive and dynamic weighting allows for results to be found in less time
- Asynchronous updates for speedy results.
Potential Minuses

- Works on a small domain
- Concept map gets cluttered quickly
- Doesn’t address any computer vision problems
- Is keyword to concept mapping the right paradigm?
- Can the automatic analytics scale?
- Authors want Automated Alert
Outline

CuZero: Frontier of Interactive Visual Search

Graph-Cut Transducers for Relevance Feedback

Non-chronological Video Synopsis and Indexing

Conclusion
An original approach to relevance feedback based on Graph-Cut
Incorporates unlabeled data
Unique vs Non-Unique categories
An example of an RF session on Corel database. First results found after submitting the top left image as a query (left) the result after 5 iterations.
Basic Approaches

- Query by Example
- Relevance Feedback
  user labels subset of images as +/- based on unknown metric
Model image set topology (include unlabeled) using a graph

Label images with binary class labels

Partition using min-cuts which is strictly equivalent to minimizing an Energy function containing:

- A fidelity term ensuring the consistency of labels of partition (provided by the user)
- A regularization term ensuring that neighboring data are likely the same label
Assumptions

- Consistent user
- Decision boundary is likely to be in low density regions of the input space
Present initial display - perhaps random - which user labels

- Train a decision algorithm

- Choose new display (techniques discussed later)
Energy Function

\[ E(S, Y) = \sum_{i=1}^{n} D_i(Y_i) + \lambda \sum_{i=1}^{n} \sum_{X_j \in N_i} V_{ij}(Y_i, Y_j) \]

Where the first term (fidelity) measures the error when mislabeling a training sample. Second term (regularizer) ensures that training samples in the neighborhood of \( X_i \) are assigned the same (or close) label.

They use a triangle kernel to measure image differences and use a Gaussian to normalize these between zero and one. Because they have this continuous distribution it can be plugged in to the Generalized Potts Model.
When labeled, Image to Sink or Source links are weighted as infinity
Display Strategies

- “Exploitation” - Select in order to refine the current estimate
  Choose unlabeled images on min-cut edges
  (efficient for single mode searches)
- “Exploration” - Find uncharted Territory
  Randomly select far from decision boundary
- “Combination” - Choose a balance
  Take a fraction of each
Exploitation vs Exploration
Evaluation

Let $K$ be the cardinality of the classes of interest. Let $Z_t$ be a random variable standing for the total number of relevant images until iteration $t$.

$$E(Z_t) = \sum_{r=1}^{K} rP(Z_t = r)$$

Also measure performance by the balanced generalization error of the classifier $f_t$ at iteration $t$.

$$\frac{1}{2} \sum_i \frac{1}{n_+} 1\{f_t(X_i) \neq \mathcal{L}(X_i) = 1\} + \frac{1}{n_-} 1\{f_t(X_i) \neq \mathcal{L}(X_i) = -1\},$$

where $n_+ = \# \{X_i, \mathcal{L}(X_i) = 1\}_{i=1}^{n}$ and $n_- = n - n_+$. 

Recall vs Iterations dependent on Neighborhood size (topology information)
Olivetti (top) and Swedish (bottom)

Far from ideal for Recall
Graphic: Display strategies dependent on class types Olivetti (top) and Swedish (bottom)
Corel

Largest disparity for Exploration, but combined shows steady growth
Olivetti (top) and Swedish (bottom)
Graph-Cut error rates are consistently best
Corel

Graph-cuts in the lead, but we stop at 30 iterations
Error rate is very choppy...
Summary

- Use an image to initialize a Query
- Choose combination Exploit/Explore images
- Create Sink/Source infinity links when labeled
- Cut and Iterate
Summary

- Use an image to initialize a Query
- Choose combination Exploit/Explore images
- Create Sink/Source infinity links when labeled
- Cut and Iterate

Questions/Issues

- Display choice is dependent on the type of data.
- Exploration is never the best strategy, maybe if data was noisier?
- 30 Iterations (too much? too little?)
Goal: Create video synopsis of movies, shortening long movies for quick viewing (http://www.vision.huji.ac.il/video-synthesis/Billiards)
Differences from previous work

- The video synopsis is itself a video, expressing the dynamics of the scene
- Reduce as much spatiotemporal redundancy as possible
- Others often fast-forward or skip frames
Recombination

![Diagram showing video synopsis and input video]

- Video synopsis
- Input video

**Recombination**
Example of splicing
Two approaches

- Region based
- Object based
Requirements

- Synopsis is substantially shorter than the original video
- Maximum “activity” (interest) from original video should appear in synopsis
- Object dynamics should be preserved
- Visible seams and fragmented objects avoided
Energy Equations

\[ E(M) = E_a(M) + \alpha E_d(M) \]

Activity of a pixel, \( \chi(x, y, t) = ||I(x, y, t) - B(x, y, t)|| \)

Activity loss, \( E_a(M) = \sum_{(x,y,t) \in I} \chi(x, y, t) - \sum_{(x,y,t) \in S} \chi(x, y, M(x, y, t)) \)

Discontinuity cost, \( E_d(M) = \sum_{(x,y,t) \in S} \sum_i ||S((x, y, t) + e_i) - I((x, y, M(x, y, t)) + e_i)||^2 \)

So across all pixels

\[ E_a(M) = \sum_{x,y} (\sum_{t=1}^K \chi(x, y, t) - \sum_{t=1}^K \chi(x, y, M(x, y) + t)) \quad \text{and} \]
\[ E_d(M) = \sum_{x,y} \sum_i \sum_{t=1}^K ||S((x, y, t) + e_i) - I((x, y, M(x, y) + t) + e_i)||^2 \]

Where \( e_i \) are the six unit vectors representing the six spatiotemporal neighbors
Ensure that the neighborhoods of A and B are similar when moving between Image and Background. This is ensured on the right by restricting consecutive synopsis pixels to come from consecutive input pixels.

Q: How are regions selected?
Construct background

- temporal median
- light to dark in 4 min chunks (surveillance cameras)

Background subtraction and min-cut isolated objects
Action tubes

[Images of airport runways and airplanes]

Yonatan Bisk
Action tubes
New Energy

New equation accounts for stitching cost

\[ E(M) = \sum_{b \in B} E_a(\hat{b}) + \sum_{b, b' \in B} (\alpha E_t(\hat{b}, \hat{b}') + \beta E_c(\hat{b}, \hat{b}')) \]

Where

- \( E_a \) is activity cost
- \( E_t \) is temporal consistency
- \( E_c \) is collision cost.
- Activity cost: penalize for object not in synopsis giving partial credit for objects cut off for lack of time
- Collision Cost: Sum of multiplied activities over shared time sequence
- Temporal consistency cost: Interaction diminishes exponentially with time
Energy Minimization

The global energy function described earlier allows us to represent as a MRF which can be optimized via Belief propagation or graph cuts. They use an unspecified "greedy algorithm."
Stroboscopic and Panoramic - Long Tubes

Aligned Space-Time Volume
Stroboscopic and Panoramic - Long Tubes
Stroboscopic and Panoramic - Obj Tracking

Coherent background and chopped up video
Endless Video

Goal is in part to be fast for querying

**Online**
- Create background by temporal medians
- Object (tube) detection and creation
- Create queue of objects
- Remove objects if queue is full

**Query stage**
- Create time lapse background
- Select tubes and compute optimal temporal arrangement
- Stitch
Removing from obj queue (Estimating obj importance)

- "importance": activity value from earlier
- "collision cost": sum of active pixels normalized and spatial distribution for obj compared for correlation
- "age": Assume density of objects in queue should decrease exponentially $N_t = K \frac{1}{\sigma} e^{-\frac{t}{\sigma}}$
Collision cost

Correlation between the two activity traces provides collision cost
Synopsis generation

- Generating background video
- Consistency cost computed for each object for each possible time
- Energy minimization determines which tubes appear and at what times
- Combine tubes with background
Time lapse background contradiction

Goal
create background of the full time of recording and background of activities

Solution
- Create Temporal histogram of activity and one of uniform time
- Interpolate to create actual video histogram
Background consistency

Want object to background consistency so new equation introduces a difference from background component to the energy function. Additionally, less than perfect segmentation so when stitching there is blending.
In Application

All the weighted components of the energy function allows users to vary variables and role of background vs scene or type of object.
Phase transition weighting

Background objects will appear and disappear for no reason
Moving objects will disappear when stopped (causes flickering)
( phase transitions should be inserted into background at original time )

![Image of a parking lot with cars and people]
Object extraction (governed by min-cut) is done in parallel and possible in hardware 3GHZ 320x240 runs at 10 fps.

Most expensive is collision cost, every relative shift between pairs of objects $K$ objects over $T$ time steps or $T \times K^2$.

**Solutions**

- Coarse intervals
- Lower resolution
- Bounding boxes
Actual times for cost computation

- 334,000 frames (24hr parking) with 262 objects becomes 450 frames in 65 seconds
- 100,000 frames (30hr airport) with 500 objects requires 80 seconds

There are $T^K$ possible temporal arrangements.

Convergence in parking example 59s and Airport 290s.

In general they throw out objects of low likelihood so airport goes from 1,917 objects to 500 from above.
Novel

- Create object tubes
- Create Median backgrounds and subtract
- Find best min collision video for a given synopsis length
**Novel**

- Create object tubes
- Create Median backgrounds and subtract
- Find best min collision video for a given synopsis length

**System changes**

- Small motions (leaves) or no motion large animals (bears) are important
- Have tubes occlude each other based on their spatial location in scene
Novel

- Create object tubes
- Create Median backgrounds and subtract
- Find best min collision video for a given synopsis length

System changes

- Small motions (leaves) or no motion large animals (bears) are important
- Have tubes occlude each other based on their spatial location in scene

User input

- Specify duration of the video synopsis and percentage of objects and try to minimize collisions
- Specify percentage of objects and penalty for collision so you optimize duration
Outline

CuZero: Frontier of Interactive Visual Search
Graph-Cut Transducers for Relevance Feedback
Non-chronological Video Synopsis and Indexing

Conclusion
<table>
<thead>
<tr>
<th>Graph-Cut</th>
<th>CuZero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query by image</td>
<td>Start with text and then allow ranking</td>
</tr>
<tr>
<td>Arbitrary set of images</td>
<td>Those with trained concept categories</td>
</tr>
</tbody>
</table>

All systems are trying to enable you to find content faster, but they work on different medium and sources.