DRESS LIKE A STAR: RETRIEVING FASHION PRODUCTS FROM VIDEOS

NOA GARCIA & GEORGE VOGIAZIS

COMPUTER VISION IN FASHION WORKSHOP
Fashion in Videos

Movies

TV shows

Online
Fashion in Videos

Sex and the City
Fashion in Videos

The Devil Wears Prada
Fashion in Videos

The Great Gastby
Fashion in Videos

Make fashion products in videos more accessible to users.
Fashion in Videos
Constraints

1. Camera view

Camera viewpoint cannot be moved to have a better view of the fashion object.
Constraints

2. User interaction

The creation of bounding boxes around the object of interest may distract users from the video.
Constraints

3. Small objects

Small, partially occluded and blurred.
Our Proposal

Instead of object recognition...
Our Proposal

Instead of object recognition... frame retrieval
Related Work

Clothing Retrieval

- Attribute classification [1]
- Domain adaptation [2]

Scene Retrieval

- Image Retrieval in Videos [3]
- Temporal tracking [4]
- Scene Descriptors [5, 6]

Our Approach: binary temporal tracking + fast indexing.
Challenges

- Average movie duration
  - 120 minutes
- Standard FPS rate
  - 24 fps
- Average frames per movie
  - 172,800 frames

With only 5 or 6 movies

More than a million frames!
Our System

Three main modules:

- Product indexing
- Training phase
- Query phase
Our System

Three main modules:

- **Product indexing**
- Training phase
- Query phase
Our System: Product indexing

Fashion items and frames related in an database.
Our System

Three main modules:

- Product indexing
- Training phase
- Query phase
Our System: Training phase

BRIEF features are more constant over time than SIFT or CNN.
Our System: Training phase

Similar frames are grouped into shots.
Our System: Training phase

- Training phase: Video frames
- Key Feature:
  - Dimension 1
  - Dimension 2
  - Dimension n
- Feature Extraction
- Feature Tracking
- Shot Detection
- Key Feature Computation
- Key Feature Indexing
- Kd-Tree Indexing
Our System

Three main modules:

- Product indexing
- Training phase
- Query phase
Our System: Query phase

Query Image

Feature Extraction
Key Feature Search

BRIEF feature

Nearest Key Features
Our System: Query phase

Query Image

Feature Extraction

Key Feature Search

Shot Search

votes

shots

key features Kf1
Our System: Query phase

Query Image → Feature Extraction → Key Feature Search → Shot Search

votes

shots

key features: Kf1, Kf2
Our System: Query phase

Query Image

Feature Extraction

Key Feature Search

Shot Search

votes

shots

key features Kf1 Kf2 Kf3
Our System: Query phase

Query Image

Feature Extraction → Key Feature Search → Shot Search → Frame Search

votes

shots

key features: Kf1 Kf2 Kf3 Kf4 Kf5 Kf6 Kf7 Kf8 Kf9

frames within the most voted shot
Our System: Query phase

Use the most similar frame to find the fashion products in the indexed product database.
Experiments - Dataset

- Webcam captures video playback.
- Frame number is used as a ground truth.

The retrieved frame should be visually similar to the annotated ground truth.
## Experiments - Retrieval Performance

Results using a single movie, 1h 49min duration

Huge gain in memory requirements with our method.

<table>
<thead>
<tr>
<th>Indexed Features</th>
<th>BF</th>
<th>KT</th>
<th>KF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85M</td>
<td>85M</td>
<td>25M</td>
<td>2M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory</th>
<th>BF</th>
<th>KT</th>
<th>KF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.53GB</td>
<td>2.53GB</td>
<td>762MB</td>
<td>61MB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>BF</th>
<th>KT</th>
<th>KF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>B = 10</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>B = 50</td>
<td>0.94</td>
<td>0.92</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>B = 100</td>
<td>0.96</td>
<td>0.93</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>B = 250</td>
<td>0.97</td>
<td>0.93</td>
<td>0.94</td>
<td></td>
</tr>
</tbody>
</table>

BF: Brute Force
KT: Kd-Tree
KF: Key Frame
Experiments - Scalability

40 movies
80 hours
7 million frames
Experiments - Scalability

Results using 40 movies

<table>
<thead>
<tr>
<th>Title</th>
<th>N. Frames</th>
<th>N. Features</th>
<th>N. Shots</th>
<th>N. Key Features</th>
<th>N. Queries</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Help</td>
<td>210387</td>
<td>101M</td>
<td>1726</td>
<td>2.2M</td>
<td>813</td>
<td>0.98</td>
</tr>
<tr>
<td>Intolerable Cruelty</td>
<td>179234</td>
<td>86M</td>
<td>1306</td>
<td>2M</td>
<td>544</td>
<td>0.97</td>
</tr>
<tr>
<td>Casablanca</td>
<td>147483</td>
<td>71M</td>
<td>881</td>
<td>1.5M</td>
<td>565</td>
<td>0.96</td>
</tr>
<tr>
<td>Witching &amp; Bitching</td>
<td>163069</td>
<td>66M</td>
<td>4193</td>
<td>0.8M</td>
<td>588</td>
<td>0.74</td>
</tr>
<tr>
<td>Pirates of the Caribbean 3</td>
<td>241127</td>
<td>108M</td>
<td>3695</td>
<td>1.7M</td>
<td>881</td>
<td>0.74</td>
</tr>
<tr>
<td>Captain Phillips</td>
<td>190496</td>
<td>59M</td>
<td>7578</td>
<td>0.6M</td>
<td>618</td>
<td>0.67</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>7M</strong></td>
<td><strong>3040M</strong></td>
<td><strong>116307</strong></td>
<td><strong>58M</strong></td>
<td><strong>25142</strong></td>
<td><strong>0.87</strong></td>
</tr>
</tbody>
</table>

Data reduction:

From 3,040M features to 58M key features.
Conclusions

- System to perform video clothing retrieval.
- It helps users to find items shown in videos.
- Based on frame retrieval and fast indexing.
- It scales well when the collection is increased.
THANK YOU!

NOA GARCIA
ASTON UNIVERSITY

CONTACT: GARCIADN@ASTON.AC.UK
GITHUB: NOAGARCIA/DRESSTAR
References


