UNDERSTANDING HUMAN-CENTRIC IMAGES
From Geometry to Fashion

Edgar Simo-Serra
Barcelona, 6th of July, 2015
MOTIVATION
MOTIVATION
MOTIVATION

Kimono

着物

Hakama

袴

Shiromuku

白無垢
MOTIVATION
MOTIVATION
MOTIVATION
MOTIVATION
MOTIVATION
MOTIVATION
Lots of information can be obtained from a single image
Location, context, roles, relationships, . . .
Prior knowledge is necessary
Must build frameworks from the ground up
This thesis is an effort towards higher level image understanding
Overview

Feature Point Descriptors
  Deformation and Light Invariant (DaLI) Descriptor
  Deep Convolutional Neural Network Descriptors

Generative 3D Human Pose Models
  Linear Latent Models
  Directed Acyclic Graphs
  Geodesic Finite Mixture Models

3D Human Pose Estimation
  3D Pose Estimation from Noisy Observations
  Joint 2D and 3D Pose Estimation

Fashion Understanding
  Semantic Segmentation of Clothing
  Modelling Fashionability

Conclusions
Overview

Prior Models

Image Description
FEATURE POINT DESCRIPTORS
Deformation and Light Invariant (DaLI) Descriptor

Deep Convolutional Neural Network Descriptors
Problem: matching points of interest under:
- Non-rigid deformations
- Photometric changes
Diffusion of heat for 3D mesh matching

Invariant to isometries

Solution is given by Heat Kernel Signature [1]

Time interval corresponds to globalness of the description

---

INVARIANCE TO DEFORMATION

Diffusion of heat for 3D mesh matching

Invariant to isometries

Solution is given by Heat Kernel Signature [1]

Time interval corresponds to globalness of the description

Embed the image as a 3D surface to apply to images

---

Diffusion of heat for 3D mesh matching

Invariant to isometries

Solution is given by Heat Kernel Signature [1]

Time interval corresponds to globalness of the description

Embed the image as a 3D surface to apply to images

Heat diffusion along the surface is used as a descriptor

\[ J. \text{Sun, M. Ovsjanikov, L. Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In Eurographics Symposium on Geometry Processing, 2009.} \]
HKS is sensitive to scale (illumination)
Use Fast Fourier Transform to gain invariance to scale [2]

---

DALI DATASET

New deformation and illumination dataset
12 objects, 4 deformation levels, 4 illumination levels
Manual annotation of correspondences

Deformation Level
Deform. Level # 0  Deform. Level # 1  Deform. Level # 2  Deform. Level # 3
New deformation and illumination dataset
12 objects, 4 deformation levels, 4 illumination levels
Manual annotation of correspondences
DaLI outperforms all, especially in illumination

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Deformation</th>
<th>Illumination</th>
<th>Deformation+ Illumination</th>
</tr>
</thead>
<tbody>
<tr>
<td>DaLI-PCA</td>
<td>67.425</td>
<td>85.122</td>
<td>68.368</td>
</tr>
<tr>
<td>DaLI</td>
<td>70.577</td>
<td>89.895</td>
<td>72.912</td>
</tr>
<tr>
<td>DAISY</td>
<td>67.373</td>
<td>75.402</td>
<td>66.197</td>
</tr>
<tr>
<td>SIFT</td>
<td>55.822</td>
<td>60.760</td>
<td>53.431</td>
</tr>
<tr>
<td>LIOP</td>
<td>58.763</td>
<td>60.014</td>
<td>52.176</td>
</tr>
<tr>
<td>Pixel Diff.</td>
<td>54.714</td>
<td>65.610</td>
<td>54.382</td>
</tr>
<tr>
<td>NCC</td>
<td>38.643</td>
<td>62.042</td>
<td>41.998</td>
</tr>
<tr>
<td>GIH</td>
<td>37.459</td>
<td>28.556</td>
<td>31.230</td>
</tr>
</tbody>
</table>
Deformation and Light Invariant (DaLI) Descriptor

Deep Convolutional Neural Network Descriptors
Learn using pairs of patches jointly
Minimize distance for "same" patches, maximize for "different" patches
Need to use all the tricks to get good performance
BEST NETWORK

Implemented in Torch7 (Lua, LeCun et al.)

INPUT: 64x64 patch

3 Convolutional layers, 46,272 parameters
1. 32x7x7 Kernel, Tanh, 2x pooling, normalization
2. 64x6x6 Kernel, Tanh, 3x pooling, normalization
3. 128x5x5 Kernel, Tanh, 4x pooling

OUTPUT: 128 dimension vector
Learn on Structure from Motion dataset
Ground truth created by using 3D structure
Sampling approach for negatives/positives
10^6 positives, 10^{12} negatives
Large amounts of mining
Essential for performance
## RESULTS

<table>
<thead>
<tr>
<th>Train</th>
<th>Test</th>
<th>SIFT</th>
<th>CNN3</th>
<th>PR AUC Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>LY+YOS</td>
<td>ND</td>
<td>0.349</td>
<td>0.667</td>
<td>91.1%</td>
</tr>
<tr>
<td>LY+ND</td>
<td>YOS</td>
<td>0.425</td>
<td>0.545</td>
<td>28.2%</td>
</tr>
<tr>
<td>YOS+ND</td>
<td>LY</td>
<td>0.226</td>
<td>0.608</td>
<td>169.0%</td>
</tr>
</tbody>
</table>

**PR curve**
- **Training LY+YOS, Test ND**
  - SIFT, CNN3, mined 1/2
  - CNN3, mined 2/2
  - CNN3, mined 4/4
  - CNN3, mined 8/8

- **Training LY+ND, Test YOS**
  - SIFT, CNN3, mined 1/2
  - CNN3, mined 2/2
  - CNN3, mined 4/4
  - CNN3, mined 8/8

- **Training YOS+ND, Test LY**
  - SIFT, CNN3, mined 1/2
  - CNN3, mined 2/2
  - CNN3, mined 4/4
  - CNN3, mined 8/8
Deformation and Light Invariant (DaLI) Descriptor

Deep Convolutional Neural Network Descriptors
CONCLUSIONS

SIFT is ubiquitous in computer vision

Better alternatives out there (DAISY, DaLI, . . .)

Alternative descriptions can be complementary

Trend to move away from hand-crafted features to learnt features continues
3D HUMAN POSE MODELS
OVERVIEW

Prior Models

Image Description
Shape Models (Hasler et al., 2009)

GPDM Models (Wang et al., 2005)
## Overview of different generative 3D human pose models

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity</th>
<th>Scales?</th>
<th>Consistent?</th>
<th>PDF?</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PGA</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GPLVM</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GPDM</td>
<td>Medium</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>hGPLVM</td>
<td>Medium</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>CRBM</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GCMFA</td>
<td>High</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>PCA</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAG</td>
<td>Medium</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GFMM</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
LINEAR LATENT MODELS (PCA)

Linear Latent Model (PCA)

Directed Acyclic Graphs (DAG)

Geodesic Finite Mixture Models (GFMM)
LINEAR LATENT MODELS (PCA)

Represent pose as a linear combination of deformation bases

\[ x = x_0 + \sum_{i=1}^{n_m} \alpha_i q_i = x_0 + Q\alpha \]

Bases found by computing SVD on the covariance of training data

\( n_m \) eigenvectors corresponding to largest eigenvalues as basis
LINEAR LATENT MODELS (PCA)

Dimension of the Latent Space $n_m$
LINEAR LATENT MODELS (PCA)

Very fast to both train and use
Linear formulation
Linear formulation
Not probabilistic
Can generate non-anthropomorphic poses
Directed Acyclic Graphs (DAG)

Linear Latent Model (PCA)

Directed Acyclic Graphs (DAG)

Geodesic Finite Mixture Models (GFMM)
Model pose with a graphical model
Probabilistically encode plausible configurations
Directed Acyclic Graph allows for dynamic programming
Model pose with a graphical model
Probabilistically encode plausible configurations
Directed Acyclic Graph allows for dynamic programming

Poses discreticized with k-means clustering
Discrete locations associated with latent states
Learnt using maximum likelihood
Efficient functions that map from latent space to pose and back
GEODESIC FINITE MIXTURE MODELS (GFMM)

Linear Latent Model (PCA)

Directed Acyclic Graphs (DAG)

Geodesic Finite Mixture Models (GFMM)
Model Probability Density Function (PDF) of data on a manifold
Fully unsupervised algorithm
Efficient implementation
One tangent space per cluster
**Geodesic distance** between two points on a manifold is the shortest distance along the manifold.
Geodesic distance between two points on a manifold is the shortest distance along the manifold.

Tangent space is a local approximation of a manifold that is a Euclidean space. Logarithm and exponential map project to and from a tangent space respectively.
Expectation-Maximization algorithm

*Minimum Message Length* used to determine number of clusters

Random initialization with large amount of clusters
Expectation-Maximization algorithm

Minimum Message Length used to determine number of clusters
Random initialization with large amount of clusters

Expectation
Data softly assigned to clusters
Statistics on tangent spaces

Expectation-Maximization algorithm

- **Minimum Message Length** used to determine number of clusters
- Random initialization with large amount of clusters

**Expectation**

- Data softly assigned to clusters

**Maximize probability with tangent spaces**

- Mean estimated on the manifold using the **geodesic mean**
- Covariance estimated on the **tangent space** in closed form
RESULTS - SYNTHETIC

Input

Cost

# Clusters

Cluster #1

Cluster #2

Cluster #3

Cluster #4

Cluster #5
Regression gives another GFMM

Scenario 1

Scenario 2

Scenario 3
Linear Latent Model (PCA)

Directed Acyclic Graphs (DAG)

Geodesic Finite Mixture Models (GFMM)
Overview of different generative 3D human pose models

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity</th>
<th>Scales?</th>
<th>Consistent?</th>
<th>PDF?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAG</td>
<td>Medium</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GFMM</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Many different ways to model the pose
Each model has different strengths/weaknesses
Exploiting known properties is beneficial
Overview of different generative 3D human pose models

<table>
<thead>
<tr>
<th>Model</th>
<th>Complexity</th>
<th>Scales?</th>
<th>Consistent?</th>
<th>PDF?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>DAG</td>
<td>Medium</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>GFMM</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Many different ways to model the pose
Each models has different strengths/weaknesses
Exploiting known properties is beneficial
Now to models in action!
3D HUMAN POSE ESTIMATION
OVERVIEW

Prior Models
Pose Estimation
Image Description
PROBLEM DEFINITION

GIVEN:

Single input image
Internal calibration A
PROBLEM DEFINITION

GIVEN:
- Single input image
- Internal calibration A

OBJECTIVE:
- Retrieve 3D pose
Single Image 3D Human Pose Estimation from Noisy Observations

A Joint Model for 2D and 3D Pose Estimation from a Single Image
3D POSE ESTIMATION FROM NOISY OBSERVATIONS

- Single Monocular 2D Image
- 2D Part Detector
- Uncertainty Propagation
- Stochastic Sampling
- Clustering
- EXPLORATION OF AMBIGUOUS HYPOTHESES
- HYPOTHESES DISAMBIGUATION
- Compute Distance Vectors
- Sort by OCSVM
- 3D Pose Estimation
3D POSE ESTIMATION FROM NOISY OBSERVATIONS

Single Monocular
2D Image

2D Part
Detector

EXPLORATION OF AMBIGUOUS HYPOTHESES

Stochastic
Sampling

Uncertainty
Propagation

Clustering

HYPOTHESES DISAMBIGUATION

Compute
Distance Vectors

Sort by
OCSVM

3D Pose
Estimation
3D POSE ESTIMATION FROM NOISY OBSERVATIONS

Single Monocular 2D Image

EXPLORATION OF AMBIGUOUS HYPOTHESES

Stochastic Sampling

Uncertainty Propagation

Clustering

HYPOTHESES DISAMBIGUATION

Compute Distance Vectors

Sort by OCSVM

3D Pose Estimation
3D POSE ESTIMATION FROM NOISY OBSERVATIONS

EXPLORATION OF AMBIGUOUS HYPOTHESES

- Single Monocular 2D Image
- 2D Part Detector
- Stochastic Sampling
- Uncertainty Propagation
- Clustering

HYPOTHESES DISAMBIGUATION

- Compute Distance Vectors
- Sort by OCSVM
- 3D Pose Estimation

Graphs and visualizations showing data points and vectors.
Projective Linear Deformation Model
Camera projection can be written as a linear equation

\[ Mx = 0 \]
Projective Linear Deformation Model
Camera projection can be written as a linear equation
Principal component analysis is also a linear equation

\[ Mx = 0 \]
\[ x = x_0 + Q\alpha \]
Projective Linear Deformation Model
Camera projection can be written as a linear equation
Principal component analysis is also a linear equation
Rank deficient system

\[
\begin{align*}
Mx &= 0 \\
x &= x_0 + Q\alpha
\end{align*}
\]

\[\implies MQ\alpha + Mx_0 = 0\]
Projective Linear Deformation Model

Camera projection can be written as a linear equation

Principal component analysis is also a linear equation

Rank deficient system

\[
\begin{align*}
    Mx & = 0 \\
    x & = x_0 + Q\alpha
\end{align*}
\]

\[\implies MQ\alpha + Mx_0 = 0\]

2D Gaussians propagated through linear system to poses

\[
\begin{align*}
    \mu_\alpha &= - (MQ)^\dagger Mx_0 \\
    \Sigma_\alpha &= \frac{\delta\alpha}{\delta\mu} \Sigma_\mu \left( \frac{\delta\alpha}{\delta\mu} \right)^T
\end{align*}
\]
## RESULTS

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S1, Walk</td>
<td></td>
<td>9.8 ± 5.3 px</td>
<td></td>
<td>24.8mm</td>
<td>46.6mm</td>
<td>805.0mm</td>
<td>55.3mm</td>
</tr>
<tr>
<td>S2, Walk</td>
<td></td>
<td>13.7 ± 5.2 px</td>
<td></td>
<td>22.1mm</td>
<td>102.4mm</td>
<td>284.5mm</td>
<td>102.4mm</td>
</tr>
<tr>
<td>S2, Jog</td>
<td></td>
<td>13.0 ± 8.8 px</td>
<td></td>
<td>15.4mm</td>
<td>68.8mm</td>
<td>157.3mm</td>
<td>68.8mm</td>
</tr>
<tr>
<td>S3, Jog</td>
<td></td>
<td>17.5 ± 13.0 px</td>
<td></td>
<td>26.3mm</td>
<td>72.6mm</td>
<td>105.3mm</td>
<td>89.5mm</td>
</tr>
</tbody>
</table>
JOINT 2D AND 3D POSE ESTIMATION

Single Image 3D Human Pose Estimation from Noisy Observations

A Joint Model for 2D and 3D Pose Estimation from a Single Image
Propose single framework for 2D and 3D
Probabilistic extendible framework
Consider image evidence to be independent for each part:

\[ p (D \mid L) = \prod_{i=1}^{N} p (d_i \mid l_i) \]
Consider image evidence to be independent for each part:

\[
p(D \mid L) = \prod_{i=1}^{N} p(d_i \mid l_i)
\]

Bayes’ rule and consider \( p(L) = p(L \mid X)p(X \mid H)p(H) \)

\[
p(X \mid D) \propto p(H)p(X \mid H) \prod_{i=1}^{N} (p(d_i \mid l_i)p(l_i \mid x_i))
\]

generative

discriminative
BAYESIAN FORMULATION

PROBABILISTIC GENERATIVE MODEL SAMPLING

DISCRIMINATION BY 2D PART DETECTORS

Best Estimation Found
## RESULTS - QUANTITATIVE

<table>
<thead>
<tr>
<th></th>
<th>Walking (A1,C1)</th>
<th>Jogging (A2,C1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
</tr>
<tr>
<td>Joint Model</td>
<td>65.1 (17.4)</td>
<td>48.6 (29.0)</td>
</tr>
<tr>
<td>Noisy Observations</td>
<td>99.6 (42.6)</td>
<td>108.3 (42.3)</td>
</tr>
<tr>
<td>[1] (tracking)</td>
<td>89.3</td>
<td>108.7</td>
</tr>
<tr>
<td>[2] (tracking)</td>
<td>-</td>
<td>107 (15)</td>
</tr>
<tr>
<td>[3] (background subtraction)</td>
<td>38.2 (21.4)</td>
<td>32.8 (23.1)</td>
</tr>
</tbody>
</table>

Single Image 3D Human Pose Estimation from Noisy Observations

A Joint Model for 2D and 3D Pose Estimation from a Single Image
Single image 3D pose estimation is an ambiguous problem  
2D evidence is very unreliable  
Strong models necessary for performance  
Joint models perform best  
  Can exploit information  
  Delay decision until the end
FASHION UNDERSTANDING
Semantic Segmentation of Clothing

Prediction
Claustrophobic Setting
User Cluster 20
Brown/Blue Jacket (2)

Recommendation
Black Casual (7)
Black Boots/Tights (4)
Black/Blue Going out (3)

Modelling Fashionability
Semantic segmentation of clothing garments
Large inter and intra class variability
Fine-grained recognition task
CONTRIBUTIONS

30% over state-of-the-art performance
Novel potentials that exploit the task
Efficient model that dresses the person

\[\text{Input} \quad \text{Truth} \quad [1] \quad \text{Ours}\]

Superpixels labels
$y_i \in \{1, \ldots, C\}$

Limb segment labels
$l_p \in \{1, \ldots, C\}$

Unary potentials
- Simple features
- Person Mask
- Clothelets
- Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs
CRF MODEL

Superpixels labels
\[ y_i \in \{1, \ldots, C\} \]

Limb segment labels
\[ l_p \in \{1, \ldots, C\} \]

Unary potentials
Simple features
Person Mask
Clothelets
Shape features

Pairwise potentials
Similarity between superpixels
Limbs

Reduce complexity
More discriminative
CRF MODEL

Superpixels labels
\( y_i \in \{1, \ldots, C\} \)

Limb segment labels
\( l_p \in \{1, \ldots, C\} \)

Unary potentials
- Simple features
- Person Mask
- Clothelets
- Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs

2D Pose Detector (Yang and Ramanan, CVPR 2011)

Problem specific
Superpixels labels $y_i \in \{1, \ldots, C\}$
Limb segment labels $l_p \in \{1, \ldots, C\}$
Unary potentials
- Simple features
  - Person Mask
  - Clothelets
- Shape features
Pairwise potentials
- Similarity between superpixels
- Limbs

Color, texture, and location histograms (Yamaguchi et al., CVPR 2012)
CRF MODEL

Superpixels labels
\( y_i \in \{1, \ldots, C\} \)

Limb segment labels
\( l_p \in \{1, \ldots, C\} \)

Unary potentials
Simple features
Person Mask
Clothelets
Shape features

Pairwise potentials
Similarity between superpixels
Limbs

Foreground/background segmentation by CPMC (Carreira and Sminchisescu, PAMI 2012)
Superpixels labels
\( y_i \in \{1, \ldots, C\} \)

Limb segment labels
\( l_p \in \{1, \ldots, C\} \)

Unary potentials
- Simple features
- Person Mask
- Clothelets
- Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs

Pose-conditioned garment likelihood
CRF MODEL

Superpixels labels
$y_i \in \{1, \ldots, C\}$

Limb segment labels
$l_p \in \{1, \ldots, C\}$

Unary potentials
- Simple features
- Person Mask
- Clothelets
  - Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs

Enriched SIFT descriptors with second order pooling (Carreira and Sminchisescu, ECCV 2012)
CRF MODEL

Superpixels labels
\( y_i \in \{1, \ldots, C\} \)

Limb segment labels
\( l_p \in \{1, \ldots, C\} \)

Unary potentials
- Simple features
- Person Mask
- Clothelets
- Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs

Shape, color, and texture similarity
(Uijlings et al., 2013)

Input
GT
Pairwise
CRF MODEL

Superpixels labels
\( y_i \in \{1, \ldots, C\} \)

Limb segment labels
\( l_p \in \{1, \ldots, C\} \)

Unary potentials
- Simple features
- Person Mask
- Clothelets
- Shape features

Pairwise potentials
- Similarity between superpixels
- Limbs

Connect superpixels to limbs using 2D pose
<table>
<thead>
<tr>
<th>Input</th>
<th>GT</th>
<th>Pose Mask</th>
<th>ShapeFeat.</th>
<th>Limbs</th>
<th>SimpleFeat.</th>
<th>FullModel</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
<td><img src="image21.png" alt="Image" /></td>
</tr>
</tbody>
</table>
RESULTS

Evaluation on the Fashionista dataset

~700 images
29 and 56 class settings

Metric is Jaccard index or Intersection over Union: \( \frac{t_p}{t_p + f_p + f_n} \)

<table>
<thead>
<tr>
<th></th>
<th>29 Classes</th>
<th>56 Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard index</td>
<td>12.32</td>
<td>20.52</td>
</tr>
</tbody>
</table>

INFERRING FASHIONABILITY AND RECOMMENDATIONS

Semantic Segmentation of Clothing

Prediction
Claustrophobic Setting
User Cluster 20
Brown/Blue Jacket (2)

Recommendation
Black Casual (7)
Black Boots/Tights (4)
Black/Blue Going out (3)

Modelling Fashionability
Large novel dataset (144,169 posts!)
Understand and model fashionability
Give fashion advice!

LOS ANGELES, CA
466 FANS
288 VOTES
62 FAVOURITES

TAGS
CHIC
EVERDAY
FALL

COLOURS
WHITE-BOOTS

NOVEMBER 10, 2014

GARMENTS
White Cheap Monday Boots
Chilli Beans Sunglasses
Missguided Romper
Daniel Wellington Watch

COMMENTS
Nice!!
Love the top!
cute
...
Large novel dataset (144,169 posts!)
Understand and model fashionability
Give fashion advice!

Prediction
Claustrophobic Setting
User Cluster 20
Brown/Blue Jacket (2)

Recommendation
Black Casual (7)
Black Boots/Tights (4)
Black/Blue Going out (3)
### FEATURES

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dim.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fans</td>
<td>1</td>
<td>Number of user’s fans.</td>
</tr>
<tr>
<td>$\Delta T$</td>
<td>1</td>
<td>Time between post creation and download.</td>
</tr>
<tr>
<td>Comments</td>
<td>5</td>
<td>Sentiment analysis [1] of comments.</td>
</tr>
<tr>
<td>Location</td>
<td>266</td>
<td>Distance from location clusters.</td>
</tr>
<tr>
<td>Personal</td>
<td>21</td>
<td>Face recognition attributes.</td>
</tr>
<tr>
<td>Style</td>
<td>20</td>
<td>Style of the photography [2].</td>
</tr>
<tr>
<td>Scene</td>
<td>397</td>
<td>Output of scene classifier trained on [3].</td>
</tr>
<tr>
<td>Tags</td>
<td>209</td>
<td>Bag-of-words with post tags.</td>
</tr>
<tr>
<td>Colours</td>
<td>604</td>
<td>Bag-of-words with colour tags.</td>
</tr>
<tr>
<td>Singles</td>
<td>121</td>
<td>Bag-of-words with split colour tags.</td>
</tr>
<tr>
<td>Garments</td>
<td>1352</td>
<td>Bag-of-words with garment tags.</td>
</tr>
</tbody>
</table>

---

Model explicitly user (u), outfit (o), setting (s), and fashionability (f). Features compressed using complementary deep networks.

CRF Models Relationships between latent states
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc.</th>
<th>Pre.</th>
<th>Rec.</th>
<th>IOU</th>
<th>L₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>29.27</td>
<td>30.42</td>
<td>28.69</td>
<td>17.36</td>
<td>1.46</td>
</tr>
<tr>
<td>Deep Net</td>
<td>30.42</td>
<td>31.11</td>
<td>30.26</td>
<td>18.41</td>
<td>1.45</td>
</tr>
<tr>
<td>Log. Reg.</td>
<td>23.92</td>
<td>22.54</td>
<td>22.99</td>
<td>12.55</td>
<td>1.91</td>
</tr>
<tr>
<td>Baseline</td>
<td>16.28</td>
<td>-</td>
<td>10.00</td>
<td>1.63</td>
<td>2.32</td>
</tr>
</tbody>
</table>

**Recommendations:** MAP estimate from conditional inference

- **Current Outfit:** Pink/Black Misc. (5)
  - Recommendations: Pastel Dress (8), Black/Blue Going out (8), Black Casual (8)

- **Current Outfit:** Pink Outfit (3)
  - Recommendations: Heels (8), Pastel Shirts/Skirts (8), Black/Gray Tights/Sweater (5)

- **Current Outfit:** Blue with Scarf (3)
  - Recommendations: Heels (8), Pastel Shirts/Skirts (8), Black Casual (8)

- **Current Outfit:** Pink/Blue Shoes/Dress Shorts (3)
  - Recommendations: Black/Gray Tights/Sweater (5), Black Casual (5), Black Boots/Tights (5)

- **Current Outfit:** Formal Blue/Brown (5)
  - Recommendations: Pastel Shirts/Skirts (9), Black/Blue Going out (8), Black Boots/Tights (8)

- **Current Outfit:** Pink/Blue Shoes/Dress Shorts (3)
  - Recommendations: Black Casual (7), Black Heavy (3), Navy and Bags (3)
Semantic Segmentation of Clothing

Prediction
- Claustrophobic Setting
- User Cluster 20
- Brown/Blue Jacket (2)

Recommendation
- Black Casual (7)
- Black Boots/Tights (4)
- Black/Blue Going out (3)

Modelling Fashionability
Fashion is very challenging!
   Hard for humans too!
Proper framework is fundamental (CRF, . . .)
Model must be tailored to the problem
Potentials must be tailored to the problem
Have to exploit as much information available
Lots of room for improvement
The new Instagram algorithm will help you dress better

Want to know what to wear, how and when? This algorithm can calculate fashionability and tell you whether you've got it, or not...

If new research from the University of Toronto is to be believed, all your common fashion problems can be solved with the help of an algorithm.

Yes, the research team have come up with a computer system to determine just how fashionable you are and help fix the problems preventing you from realatorial success.
NEW ALGORITHM FOR INSTAGRAM WILL TELL YOU HOW TO DRESS BETTER

For maximum likes.

If you're anything like us, your Instagram feed is full of carefully framed selfies and snaps of your latest #ootd. (Because, what else?) The app was pretty much made for
Inventan un software que mide tu nivel de estilo

Más sincero que la opinión de tu madre y de tu mejor amiga juntas, este programa puede valorar tu look, criticarlo y ofrecer consejos para mejorarlo.

Beatriz Caballero — @bethferrier — Las ocasiones importantes, requieren looks importantes y su aprobación pasa siempre por el visto de la amiga o persona de confianza de turno. Que levante la mano la que no se haya hecho una foto delante del espejo con el modelito escogido o en el probador de su tienda favorita y la haya enviado a ese grupo de whatsapp, esperando impacientemente el veredicto.

Pero este método, no nos engañemos, tiene fisuras... ¿Es esa amiga realmente sincera? ¿El gusto de tu amiga es el que más prevalece sobre tu propio criterio? ¿Ese grupo de whatsapp es realmente cool para valorar si ese es tu mejor outfit para acudir a tu cita o a tu nuevo puesto de trabajo? Y, en tu modo realmente...
CONCLUSIONS

FASHION / 23 JUNE 15 / by KATIE COLLINS

Are you ever worried your friends aren’t being honest with you about your taste in fashion? Well now you can turn to a computer in order to have your suspicions confirmed, or not, as the case may be. Don’t be downcast if it turns out your outfits are not de rigeur, however -- the software can also help you plan a Cher from Clueless-style makeover.

Researchers at the University of Toronto have developed an algorithm that uses a combination of computer vision and
CONCLUSIONS
Feature Point Descriptors
- Deformation and Light Invariant (DaLi) Descriptor
- Deep Convolutional Neural Network Descriptors

Generative 3D Human Pose Models
- Linear Latent Models
- Directed Acyclic Graphs
- Geodesic Finite Mixture Models

3D Human Pose Estimation
- 3D Pose Estimation from Noisy Observations
- Joint 2D and 3D Pose Estimation

Fashion Understanding
- Semantic Segmentation of Clothing
- Modelling Fashionability


CONCLUSIONS

Low to high level overview of human-centric computer vision.
Computer Vision and Machine Learning are highly intertwined.
Must stay up to date and exploit existing tools.
Code is available for most projects [1]

FUTURE DIRECTIONS

Holistic models - multiple tasks at once
More real world applications
More features from state of the art (deep networks)
FUTURE DIRECTIONS

Holistic models - multiple tasks at once
More real world applications
More features from state of the art (deep networks)
Thanks to my directors, collaborators, friends, family, supporters, and most importantly...
Thanks to my directors, collaborators, friends, family, supporters, and most importantly...
Questions?

http://www.iri.upc.edu/people/esimo/