Recognizing Proxemics in Personal Photos

Yi Yang Simon Baker Anitha Kannan Deva Ramanan

UC Irvine Microsoft Research UC Irvine

Introduction

(a) Hand-hand (b) Shoulder-shoulder (c) Hand-shoulder

(d) Hand-elbow (e) Elbow-shoulder (f) Hand-torso

(g) (h) (i) (j)

Proxemics[1] is the study of how people interact. We present a computational approach of visual proxemics by labeling each pair of people with a set of **touch codes**, defined as the pairs of body parts (each element of the pair comes from a different person) that are in physical contact.

- (a)-(f) Six specific touch codes that we study in this paper.
- (g)-(j) Illustration of wide variation in appearance for hand-hand proxemic.
- (g) Also illustrates that multiple touch codes may appear at the same time.

Dataset

(a) Image Statistics

No. Images	No. People	No. People Pairs
589	1207	1332

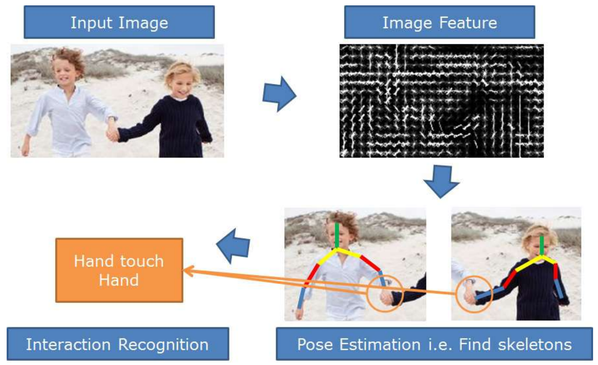
(b) Touch Code Statistics

Hand-hand	Hand-shoul	Shoul-shoul	Hand-elbow	Elbow-shoul	Hand-torso
340	180	210	96	106	57
25.5%	13.5 %	15.8 %	7.2%	8.0%	4.3%

(c) Co-occurrence Statistics

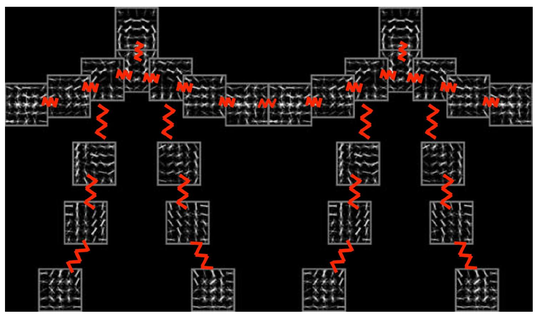
0 Codes	1 Code	2 Codes	3+ Codes
531	626	162	13

Baseline - Sequential Method



A baseline approach would be to first perform pose estimation[2] and then detect touch codes based on the estimated joint locations. However, this sequential approach does **not** perform well because pose estimation step is too unreliable for images of interacting people due to occlusion and part ambiguity.

Our Model - Joint Method



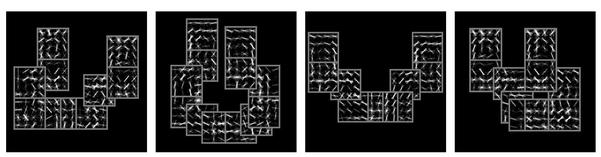
Our model for hand-hand proxemic is a pictorial structure consisting of two people plus a spring connecting their hands, shown as HOG template.

We augement the standard pictorial structure model[3]:

$$S(I, L) = \Sigma_{i \in V}\ \alpha_i \cdot \phi(I, l_i) + \Sigma_{ij \in E}\ \beta_{ij} \cdot \psi(l_i, l_j)$$

- I : image window
- l_i : the pixel location of part i
- $\phi(I, l_i)$: local appearance feature (e.g. HOG) extracted from location l_i
- $\psi(l_i, l_j)$: spatial feature extracted from the relative location l_i w.r.t. l_j
- α_i : local appearance template for part i
- β_{ij} : spatial pairwise spring parameter for part i and j

Submixture Model



(a) Left left (b) Left right (c) Right left (d) Right right

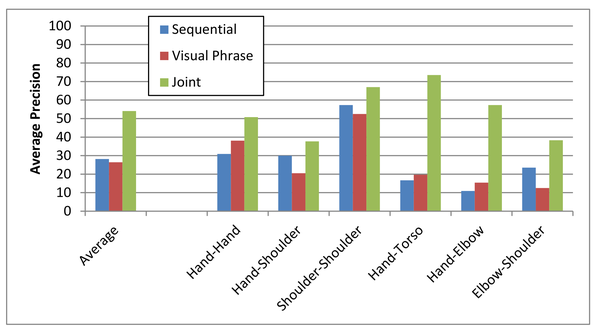
As each person has two arms, we use four sub-models to capture the different hand-hand appearances. The maximum likely one is taken during inference.

Model Visualization and Pose Estimation Results

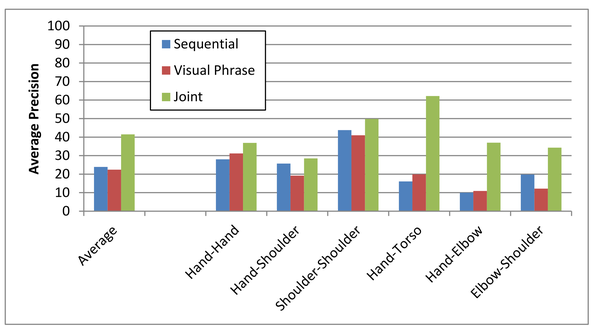
(a) Hand-hand (b) Shoulder-shoulder (a) Hand-shoulder (d) Hand-elbow (c) Elbow-shoulder (e) Hand-torso

- (1st row) Illustration of tree-structure of our proxemic-specific model. As it is not important to consider the legs and other arms/torso parts to predict the proxemics, we crop out those regions and build a chain connecting from one person's head to the other person's head through the touching body parts.
- (2st row) Sample results for pose estimation using **sequential** model which independently estimates poses of each person.
- (3rd row) Sample results for pose estimation using joint model but **without key spring** where the spring connecting the two bodies is removed.
- (4th row) Sample results for pose estimation using our **joint** model which produces more reliable pose estimates because it better models occlusions and spatial constraints specific to each touch code.

Proxemics Classification Results



(a) Using ground truth head locations



(b) Using face detection to identifying head locations

Comparison between our proposed **joint** algorithm, the **sequential** algorithm, and the **visual phrase** algorithm[4]. In (a) we use the ground-truth head positions. In (b) we use the faces obtained using a face detector. Our algorithm gives a very significant improvement in average precision in both cases.

References

[1] E. Hall. A system for the notation of proxemic behavior. American anthropologist, 65(5), 1963.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. IJCV, 61(1), 2005.

[3] Y. Yang and D. Ramanan. Articulated pose estimation using flexible mixtures of parts. In CVPR, 2011.

[4] A. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, 2011.