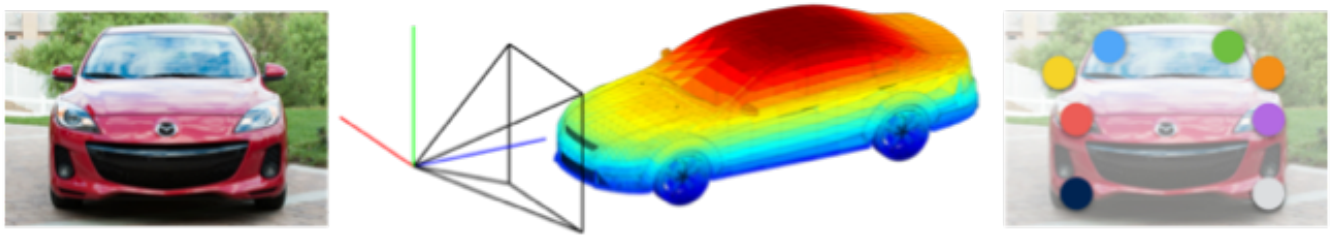
Estimating pose and locating keypoints

Jitendra Malik

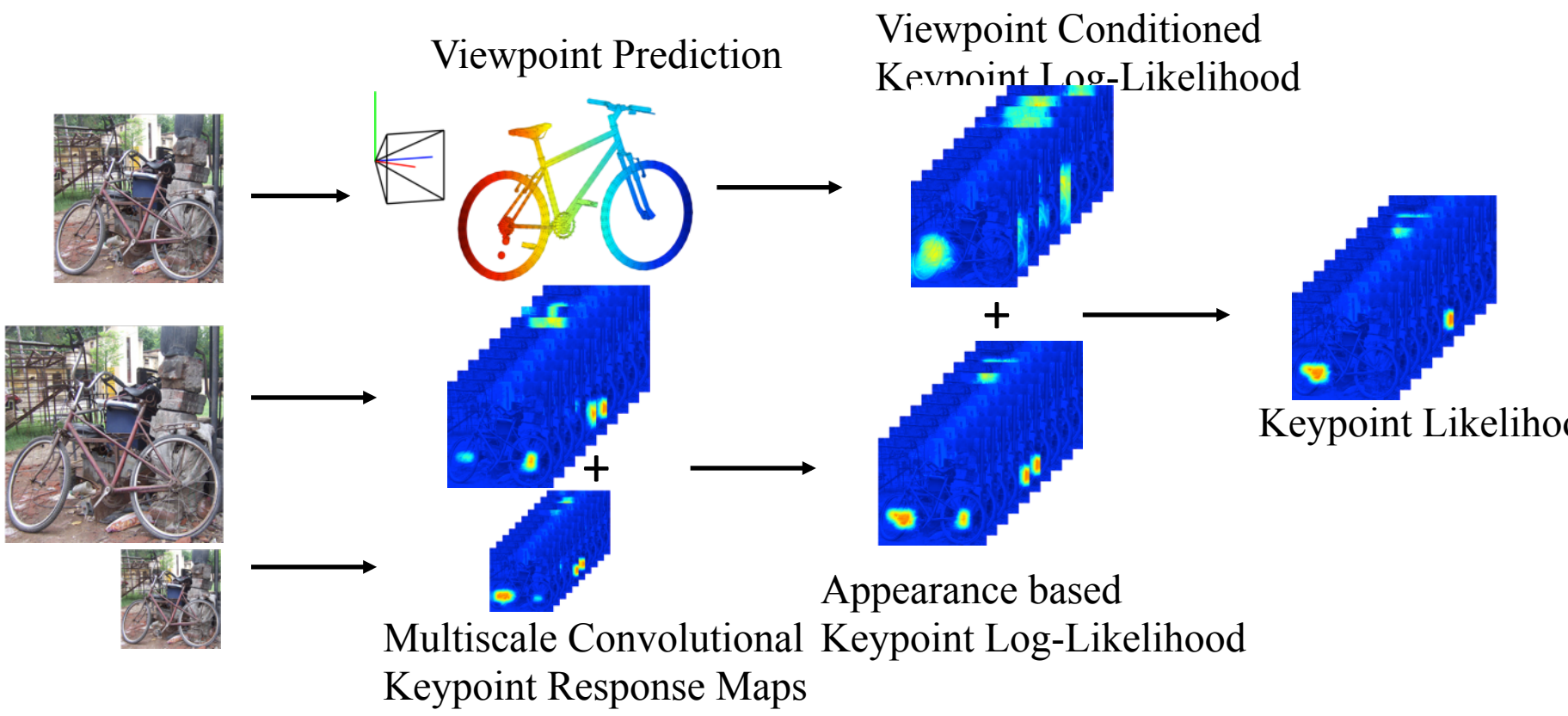
Viewpoints and Keypoints

- Shubham Tulsiani, Jitendra Malik



We characterize the problem of pose estimation for rigid objects in terms of determining viewpoint to explain coarse pose and keypoint prediction to capture the finer details. We address both these tasks in two different settings - the constrained setting with known bounding boxes and the more challenging detection setting where the aim is to simultaneously detect and correctly estimate pose of objects. We present Convolutional Neural Network based architectures for these and demonstrate that leveraging viewpoint estimates can substantially improve local appearance based keypoint predictions. In addition to achieving significant improvements over state-of-the-art in the above tasks, we analyze the error modes and effect of object characteristics on performance to guide future efforts towards this goal.

Overview

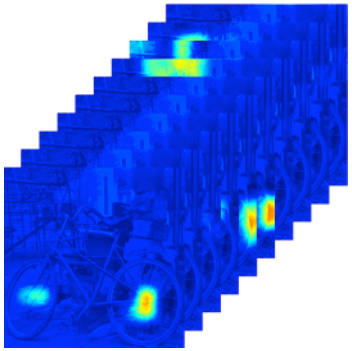

Viewpoint Prediction
Viewpoint Conditioned
Keypoint Log-Likelihood
+
+
Keypoint Likelihoo
Multiscale Convolutional
Keypoint Response Maps
Appearance based
Keypoint Log-Likelihood

Viewpoint Prediction

- Viewpoint is described by elevation, azimuth and cyclorotation
- For each angle, the problem is treated as a classification among fixed bins
- We train a VGG based CNN where fc-8 does predicts the bin for each euler angle
- Some tricks used to jointly train a network for all classes (loss computed only on class specific fc8 units)

Keypoint Prediction : Convolutional Responses

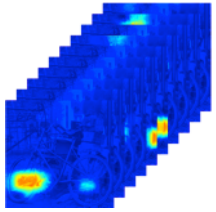
416 X 416 image



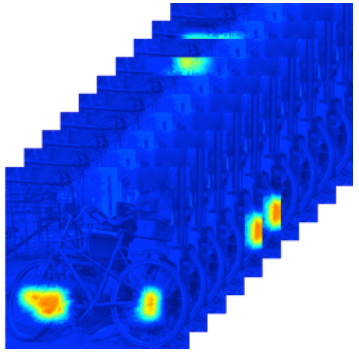
12 X 12 response map

+

224 X 224 image



6 X 6 response map

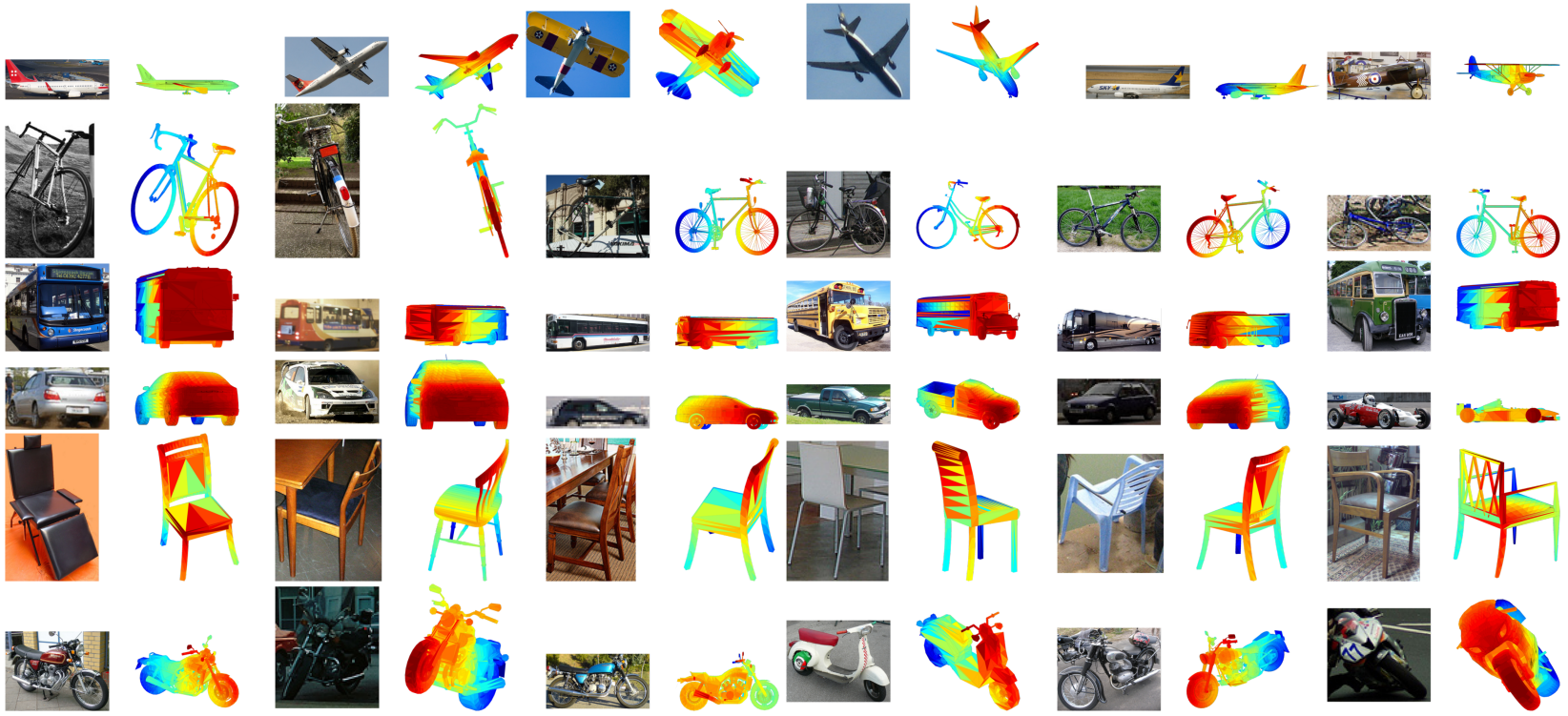


12 X 12 response map

Keypoint Prediction : Viewpoint Based Prior

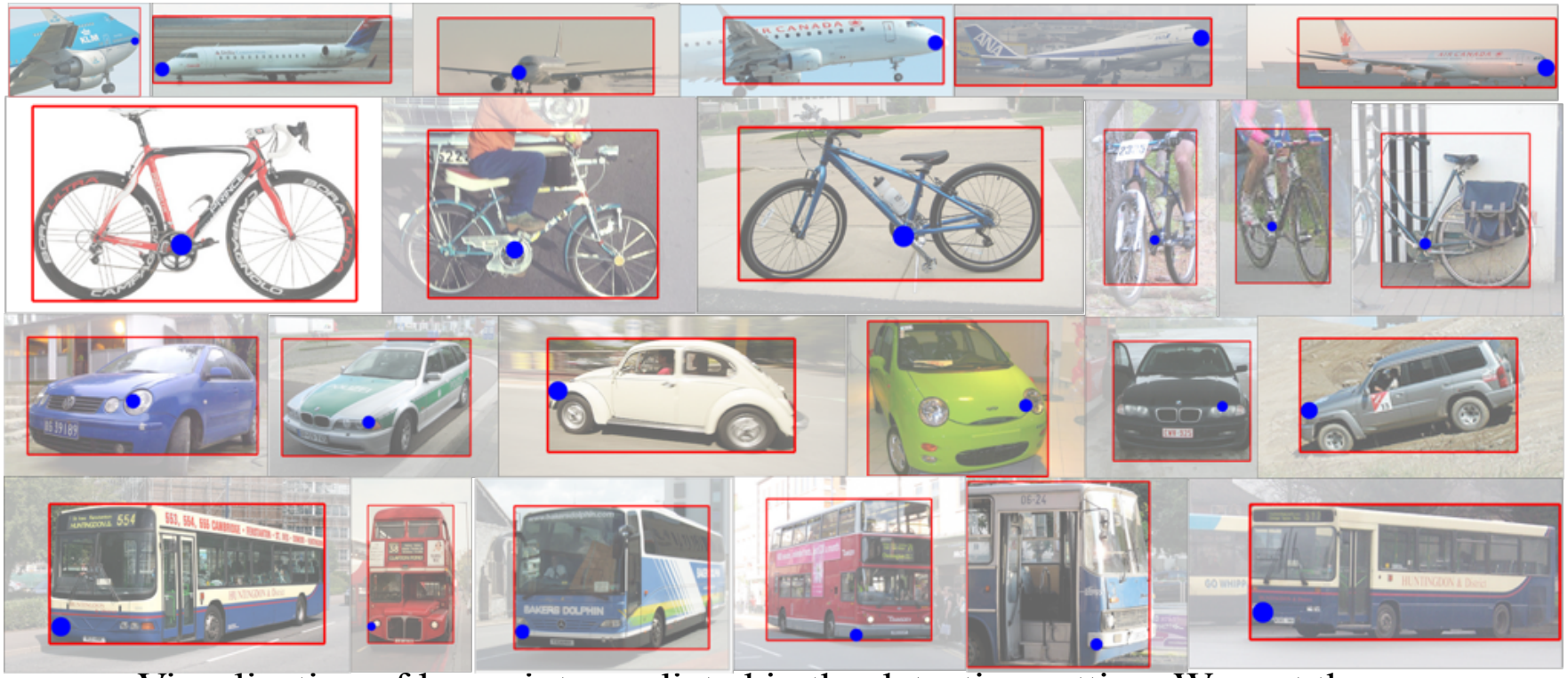
The viewpoint based location prior is computed using keypoint locations of training instances with similar views

Examples : Viewpoint Prediction



The columns show 15th, 30th, 45th, 60th, 75th and 90th percentile instances in terms of the error.

Examples : Keypoint Prediction



Visualization of keypoints predicted in the detection setting. We sort the keypoints detections by their prediction score and visualize every 15th detection for 'Nosetip' of aeroplanes, 'Left Headlight' of cars, 'Crankcentre' of bicycles and 'Left Base' of buses.

Results : Viewpoint Prediction

	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
$Acc_{\frac{\pi}{6}}$ (Pool5-TNet)	0.27	0.18	0.36	0.81	0.71	0.36	0.52	0.52	0.38	0.67	0.7	0.71	0.52
$Acc_{\frac{\pi}{6}}$ (fc7-TNet)	0.5	0.44	0.39	0.88	0.81	0.7	0.39	0.38	0.48	0.44	0.78	0.65	0.57
$Acc_{\frac{\pi}{6}}$ (ours-TNet)	0.78	0.74	0.49	**0.93**	0.94	**0.90**	0.65	**0.67**	0.83	0.67	0.79	0.76	0.76
$Acc_{\frac{\pi}{6}}$ (ours-ONet)	**0.81**	**0.77**	**0.59**	**0.93**	**0.98**	0.89	**0.80**	0.62	**0.88**	**0.82**	**0.80**	**0.80**	**0.81**
$MedErr$ (Pool5-TNet)	42.6	52.3	46.3	18.5	17.5	45.6	28.6	27.7	37	25.9	20.6	21.5	32
$MedErr$(fc7-TNet)	29.8	40.3	49.5	13.5	7.6	13.6	45.5	38.7	31.4	38.5	9.9	22.6	28.4
$MedErr$(ours-TNet)	14.7	18.6	31.2	13.5	6.3	**8.8**	17.7	17.4	17.6	15.1	8.9	17.8	15.6
$MedErr$(ours-ONet)	**13.8**	**17.7**	**21.3**	**12.9**	**5.8**	9.1	**14.8**	**15.2**	**14.7**	**13.7**	**8.7**	**15.4**	**13.6**

Viewpoint Prediction (known bounding box)

	AVP				$AVP_{\frac{\pi}{6}}$	$ARP_{\frac{\pi}{6}}$
Number of bins	**4**	**8**	**16**	**24**	**-**	**-**
Xiang *et al*. [37]	19.5	18.7	15.6	12.1	-	-
Pepik *et al*. [30]	23.8	21.5	17.3	13.6	-	-
Ghodrati *et al*. [9]	24.1	22.3	17.3	13.7	-	-
ours	**49.1**	**44.5**	**36.0**	**31.1**	50.7	46.5

Viewpoint Prediction (detection setting)

Results : Keypoint Prediction

PCK[$\alpha = 0.1$]	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
Long *et al.* [25]	53.7	60.9	33.8	72.9	70.4	55.7	18.5	22.9	52.9	38.3	53.3	49.2	48.5
conv6 (coarse scale)	51.4	62.4	37.8	65.1	60.1	59.9	34.8	31.8	53.6	44	52.3	41.1	49.5
conv12 (fine scale)	54.9	66.8	32.6	60.2	80.5	59.3	35.1	37.8	58	41.6	59.3	53.8	53.3
conv6+conv12	61.9	74.6	43.6	72.8	84.3	70.0	45.0	44.8	66.7	51.2	66.8	56.8	61.5
conv6+conv12+pLikelihood	**66.0**	**77.8**	**52.1**	**83.8**	**88.7**	**81.3**	**65.0**	**47.3**	**68.3**	**58.8**	**72.0**	**65.1**	**68.8**

Keypoint Localization (known bounding box)

APK[$\alpha = 0.1$]	aero	bike	boat	bottle	bus	car	chair	table	mbike	sofa	train	tv	mean
conv6+conv12	41.9	47.1	15.4	29.0	58.2	37.1	11.2	8.1	40.7	25.0	36.9	25.5	31.3
conv6+conv12+pLikelihood	**44.9**	**48.3**	**17.0**	**30.0**	**60.8**	**40.7**	**14.6**	**8.6**	**42.8**	**25.7**	**38.3**	**26.2**	**33.2**

Keypoint Detection (no bounding box)

Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation

Jonathan Tompson, Arjun Jain, Yann LeCun, Christoph Bregler
New York University
{tompson, ajain, yann, bregler}@cs.nyu.edu

Abstract

This paper proposes a new hybrid architecture that consists of a deep Convolutional Network and a Markov Random Field. We show how this architecture is successfully applied to the challenging problem of articulated human pose estimation in monocular images. The architecture can exploit structural domain constraints such as geometric relationships between body joint locations. We show that joint training of these two model paradigms improves performance and allows us to significantly outperform existing state-of-the-art techniques.

Key ideas

- Convolutional net based part detector predicts heat maps for keypoints.
- Use these as unary potentials for a MRF with binary potentials that constrain joint inter-connectivity and global pose consistency (e.g. a face detector peak should not be very far from a shoulder detector peak)

3.1 Convolutional Network Part-Detector

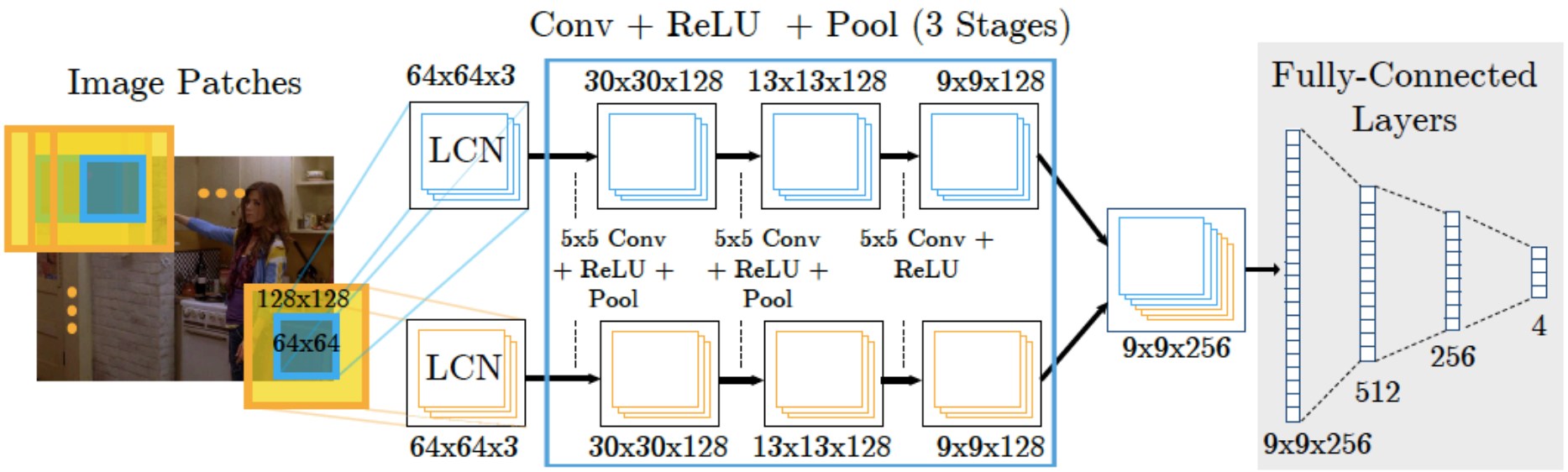


Figure 1: Multi-Resolution Sliding-Window With Overlapping Receptive Fields

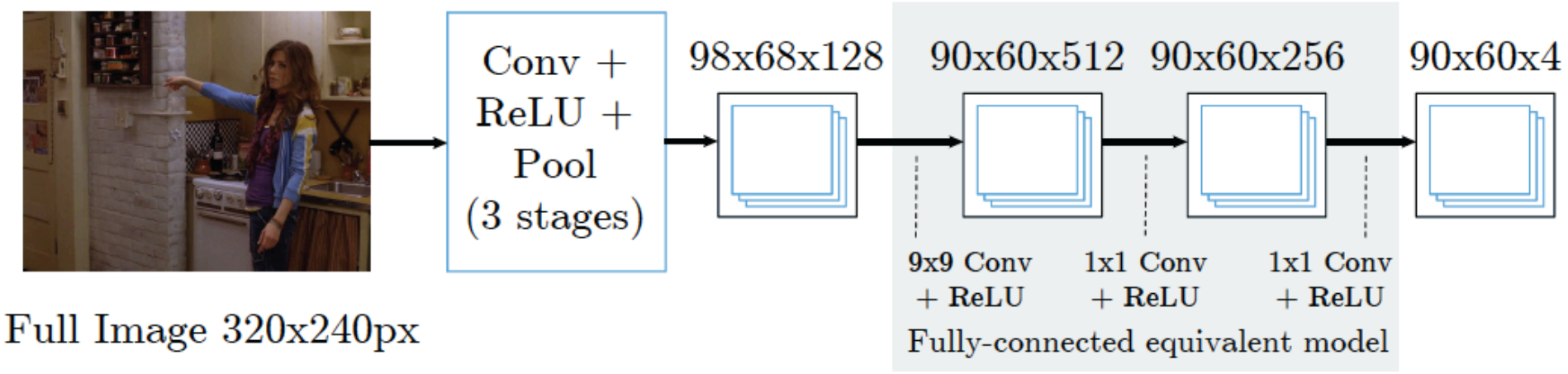


Figure 2: Efficient Sliding Window Model with Single Receptive Field

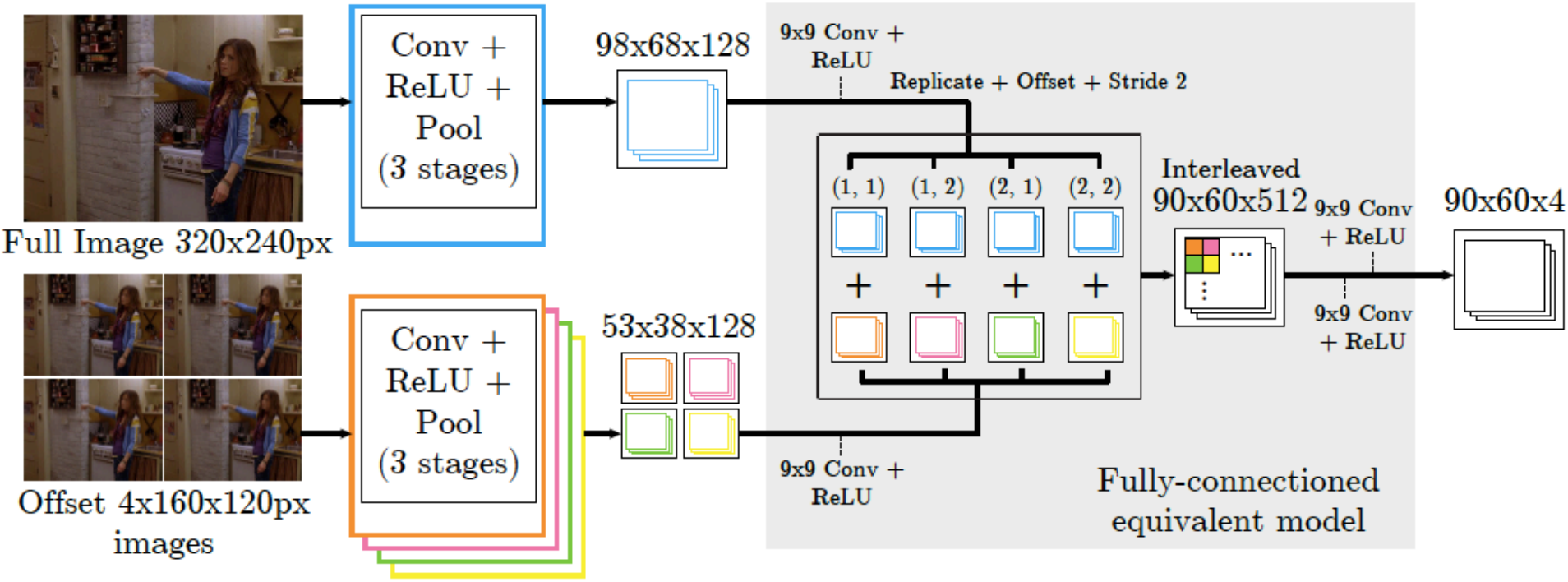


Figure 3: Efficient Sliding Window Model with Overlapping Receptive Fields

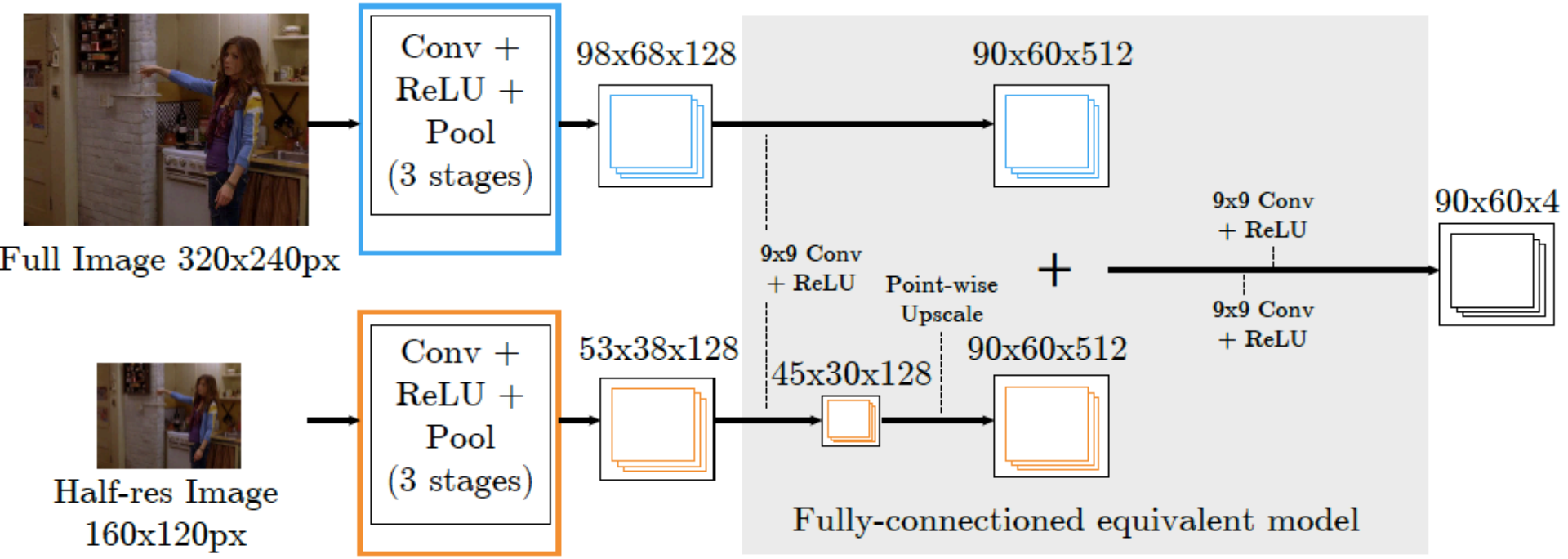


Figure 4: Approximation of Fig 3

Convolutional priors

given that body part B is located at the center pixel, the convolution prior $P_{A|B}(i,j)$ is the likelihood of the body part A occurring in pixel location (i,j). For a body part A, we calculate the final marginal likelihood $\bar{p}_A$ as:

$$\bar{p}_A = \frac{1}{Z} \prod_{v \in V} \left(p_{A|v} * p_v + b_{v \to A} \right) \tag{1}$$

where v is the joint location, $p_{A|v}$ is the conditional prior described above, $b_{v \to a}$ is a bias term used to describe the background probability for the message from joint v to A

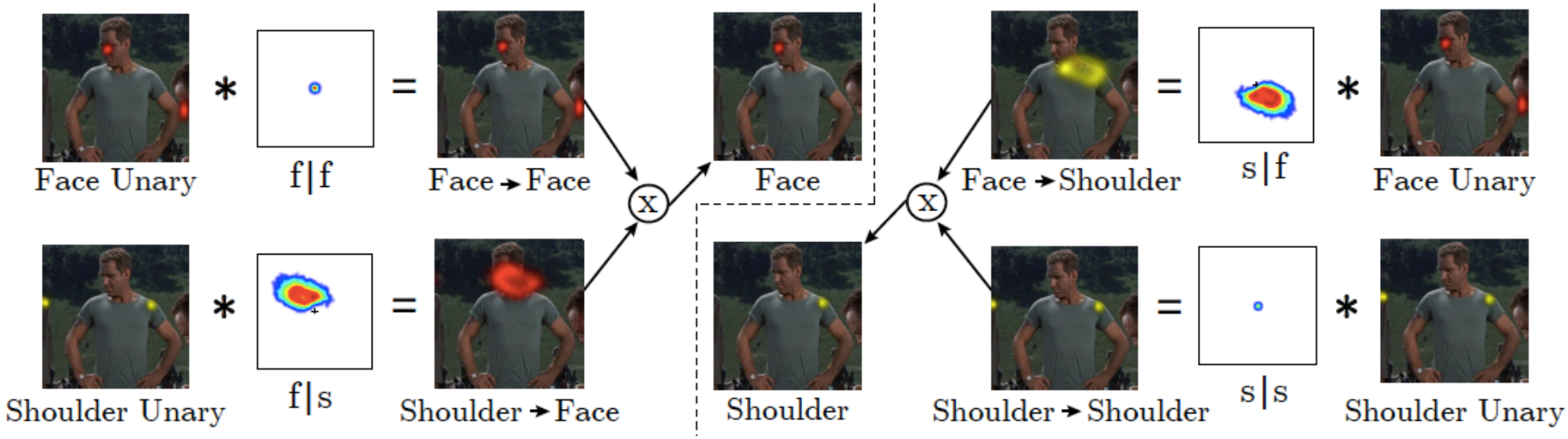


Figure 5: Didactic Example of Message Passing Between the Face and Shoulder Joints

$$\bar{e}_A = \exp\left(\sum_{v \in V}\left[\log\left(\text{SoftPlus}\left(e_{A|v}\right) * \text{ReLU}\left(e_v\right) + \text{SoftPlus}\left(b_{v \to A}\right)\right)\right]\right) \tag{2}$$

$$\text{where: } \text{SoftPlus}\left(x\right) = {}^{1}/_{\beta}\log\left(1 + \exp\left(\beta x\right)\right), {}^{1}/_{2} \leq \beta \leq 2$$

$$\text{ReLU}\left(x\right) = \max\left(x, \epsilon\right), 0 < \epsilon \leq 0.01$$

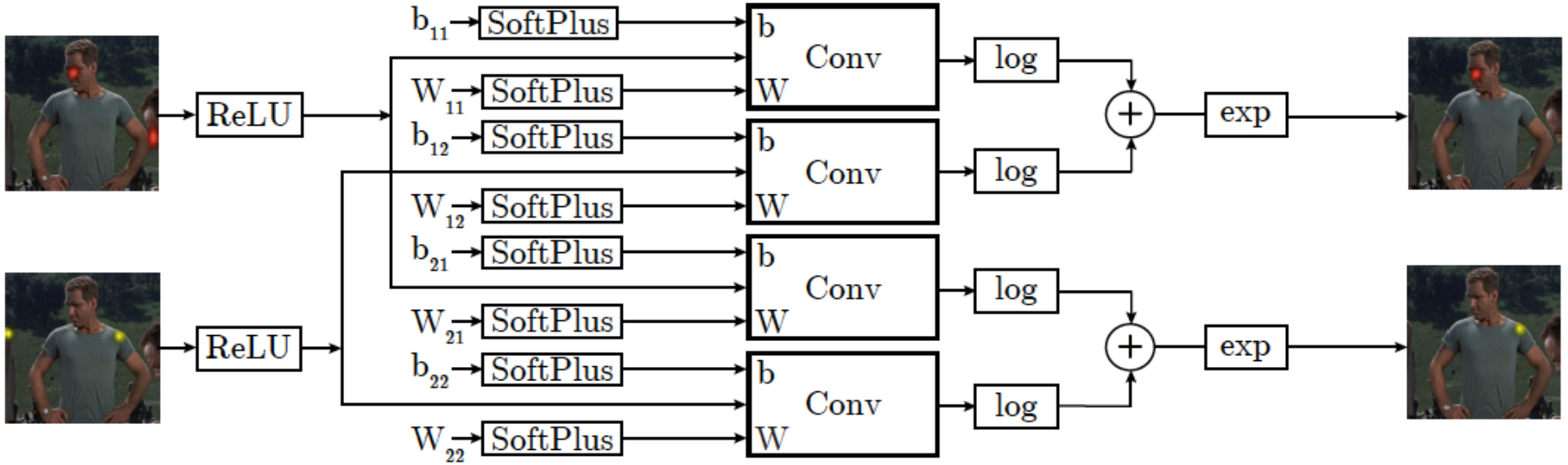
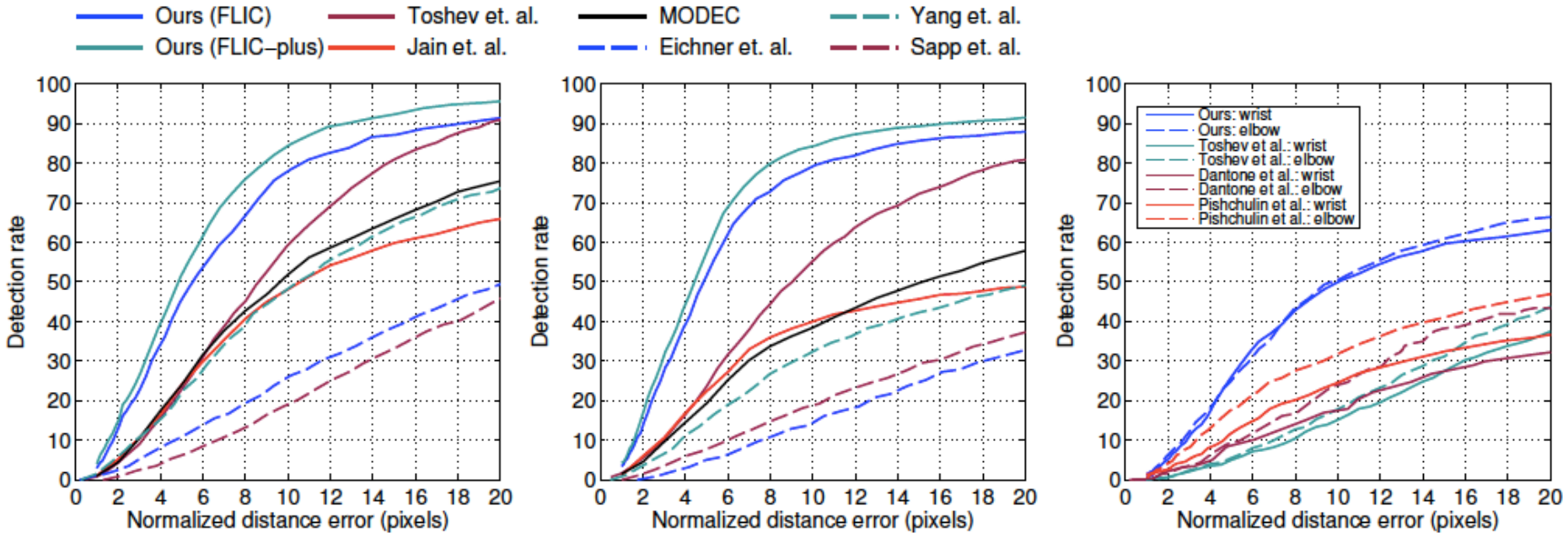
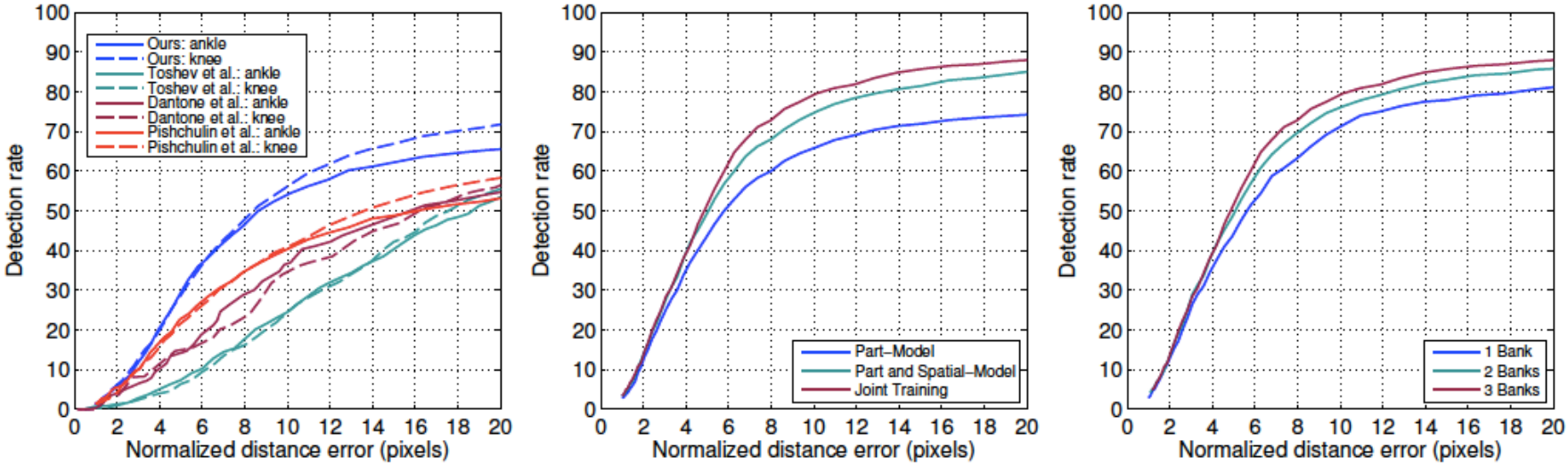


Figure 6: Single Round Message Passing Network



(a) FLIC: Elbow (b) FLIC: Wrist (c) LSP: Wrist and Elbow

Figure 7: Model Performance



(a) LSP: Ankle and Knee (b) FLIC: Wrist (c) FLIC: Wrist

Figure 8: (a) Model Performance (b) With and Without Spatial-Model (c) Part-Detector Performance Vs Number of Resolution Banks (FLIC subset)

International Journal of Computer Vision, 5:2, 195–212 (1990)

Recognizing Solid Objects by Alignment with an Image

DANIEL P. HUTTENLOCHER
Computer Science Department, Cornell University, 4130 Upson Hall, Ithaca, NY 14850

SHIMON ULLMAN
Department of Brain and Cognitive Science, and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

In this paper we consider the problem of recognizing solid objects from a single two-dimensional image of a three-dimensional scene. We develop a new method for computing a transformation from a three-dimensional model coordinate frame to the two-dimensional image coordinate frame, using three pairs of model and image points. We show that this transformation always exists for three noncollinear points, and is unique up to a reflective ambiguity. The solution method is closed-form and only involves second-order equations. We have implemented a recognition system that uses this transformation method to determine possible *alignments* of a model with an image. Each of these hypothesized matches is verified by comparing the entire edge contours of the aligned object with the image edges. Using the entire edge contours for verification, rather than a few local feature points, reduces the chance of finding false matches. The system has been tested on partly occluded objects in highly cluttered scenes.

Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image

Camillo J. Taylor

GRASP Laboratory, CIS Department, University of Pennsylvania, 3401 Walnut Street, Rm 335C, Philadelphia, Pennsylvania 19104-6228

E-mail: cjtaylor@central.cis.upenn.edu

Received July 19, 1999; accepted September 11, 2000

This paper investigates the problem of recovering information about the configuration of an articulated object, such as a human figure, from point correspondences in a single image. Unlike previous approaches, the proposed reconstruction method does not assume that the imagery was acquired with a calibrated camera. An analysis is presented which demonstrates that there is a family of solutions to this reconstruction problem parameterized by a single variable. A simple and effective algorithm is proposed for recovering the entire set of solutions by considering the foreshortening of the segments of the model in the image. Results obtained by applying this algorithm to real images are presented.

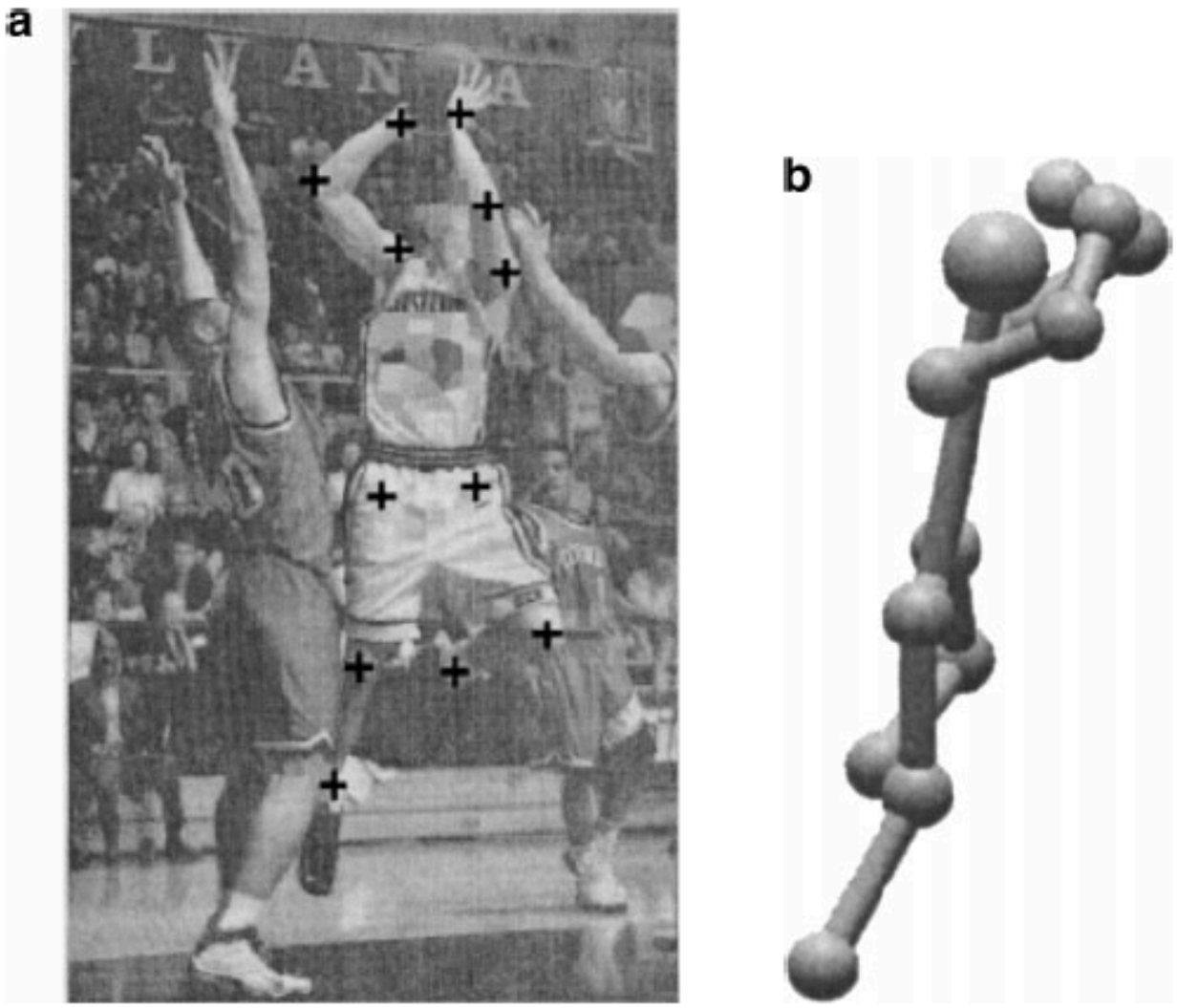
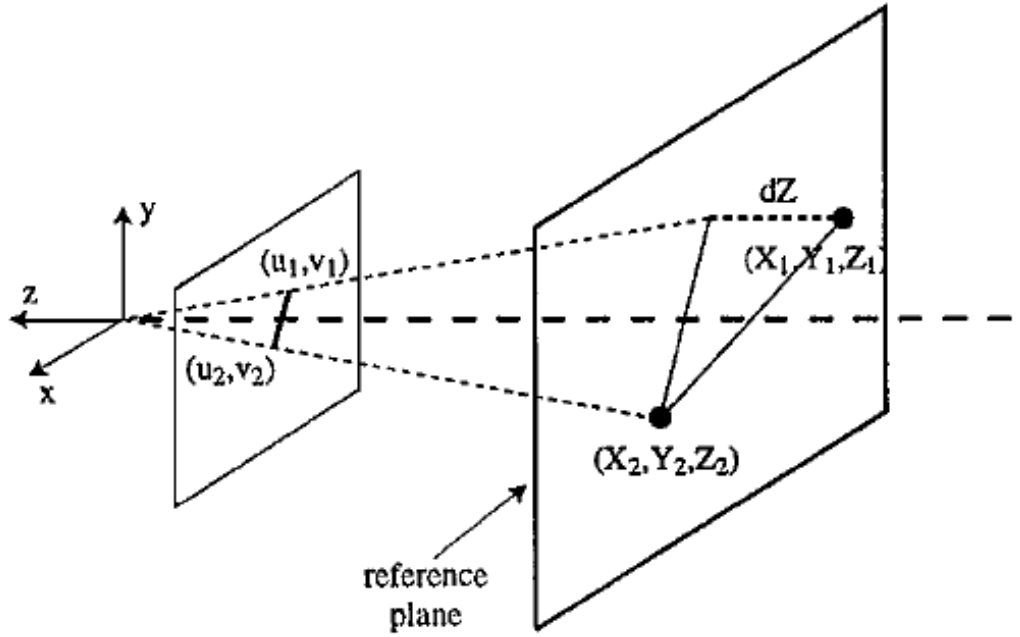


FIG. 1. (a) An image containing a figure to be recovered. The 12 crosses represent the estimated locations of the joints which are passed to the reconstruction procedure. (b) The recovered 3D model viewed from a novel vantage point.

Key observation: Lengths of various body segments are known



$$\begin{pmatrix} u \\ v \end{pmatrix} = s \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

The projection of a line segment onto an image under scaled orthographic projection

$$l^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2$$

$$(u_1 - u_2) = s(X_1 - X_2)$$

$$(v_1 - v_2) = s(Y_1 - Y_2)$$

$$dZ = (Z_1 - Z_2)$$

$$\Rightarrow dZ = \sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)/s^2}$$

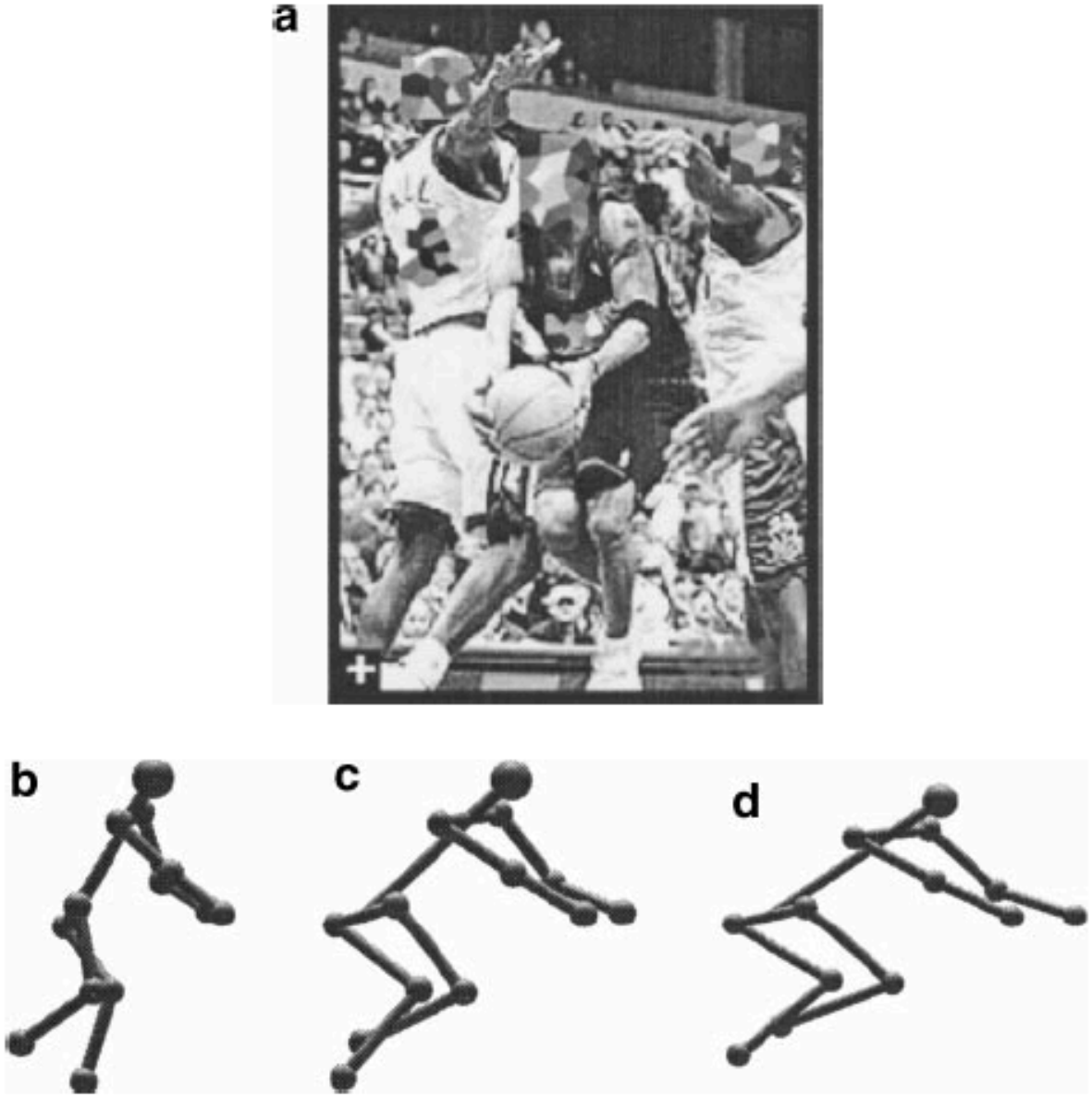


FIG. 5. This figure indicates how the reconstructions computed from the point correspondences obtained from the image in (a) vary as a function of the scale parameter s. The reconstructions shown in (b), (c), and (d) correspond to scale factor values of 2.3569, 2.9461, and 3.5353, respectively.

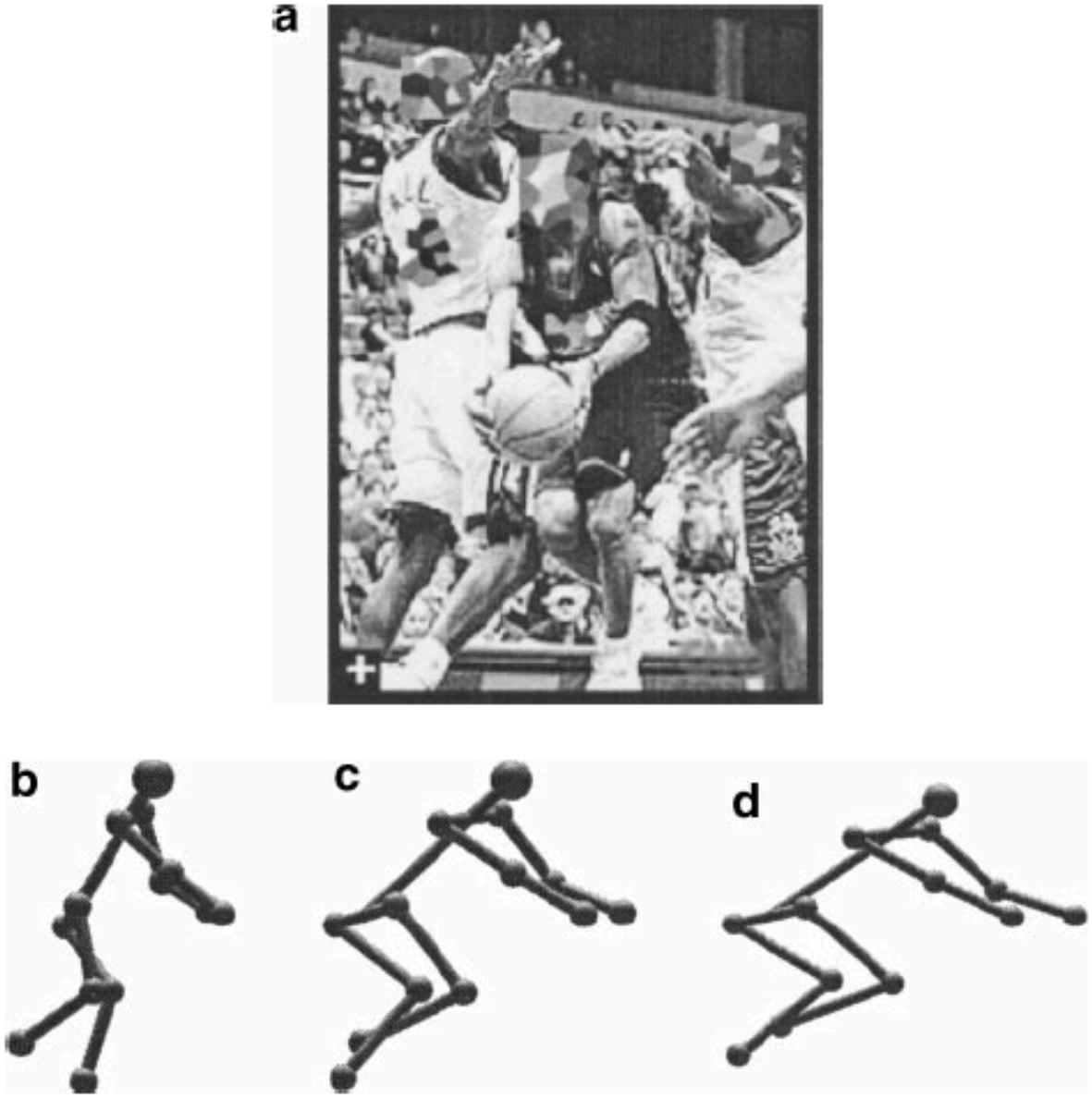


FIG. 5. This figure indicates how the reconstructions computed from the point correspondences obtained from the image in (a) vary as a function of the scale parameter s. The reconstructions shown in (b), (c), and (d) correspond to scale factor values of 2.3569, 2.9461, and 3.5353, respectively.