Modeling and recognizing interactions between people, objects and scenes

Vincent Delaitre

Thèse

présentée à

l'École Normale Supérieure
École Doctorale de sciences
mathématiques de Paris-Centre

par

Vincent Delaitre

pour obtenir

le grade de docteur en sciences
spécialité : informatique

Modeling and Recognizing Interactions between People, Objects and Scenes

Directeurs de thèse : Ivan Laptev et Josef Sivic

Soutenue le 7 avril 2015, devant la commission d'examen formée de :

Vittorio Ferrari	Professeur, University of Edinburgh	Rapporteur
Derek Hoiem	Professeur, University of Illinois	Rapporteur
Patrick Perez	Chercheur Associé, Technicolor	Examinateur
Jean Ponce	Professeur, École Normale Supérieure	Examinateur
Cordelia Schmid	Directrice de Recherche, Inria Grenoble	Examinateur
Ivan Laptev	Directeur de Recherche, Inria Rocquencourt	Dir. de thèse
Josef Sivic	Chargé de Recherche, Inria Rocquencourt	Dir. de thèse

Résumé

Nous nous intéressons dans cette thèse à la modélisation des interactions entre personnes, objets et scènes. Nous montrons l'intérêt de combiner ces trois sources d'information pour améliorer la classification d'action et la compréhension automatique des scènes. Dans la première partie, nous cherchons à exploiter le contexte fourni par les objets et la scène pour améliorer la classification des actions humaines dans les photographies. Nous explorons différentes variantes du modèle dit de "bag-of-features" et proposons une méthode tirant avantage du contexte scénique. Nous proposons ensuite un nouveau modèle exploitant les objets pour la classification d'action basé sur des paires de détecteurs de parties du corps et/ou d'objet. Nous évaluons ces méthodes sur notre base de données d'images nouvellement collectée ainsi que sur trois autres jeux de données pour la classification d'action et obtenons des résultats proches de l'état de l'art.

Dans la seconde partie de cette thèse, nous nous attaquons au problème inverse et cherchons à utiliser l'information contextuelle fournie par les personnes pour aider à la localisation des objets et à la compréhension des scènes. Nous collectons une nouvelle base de données de time-lapses comportant de nombreuses interactions entre personnes, objets et scènes. Nous développons une approche permettant de décrire une zone de l'image par la distribution des poses des personnes qui interagissent avec et nous utilisons cette représentation pour améliorer la localisation d'objets. De plus, nous démontrons qu'utiliser des informations provenant des personnes détectées peut améliorer plusieurs étapes de l'algorithme utilisé pour la compréhension des scènes d'intérieur. Pour finir, nous proposons des annotations 3D de notre base de time-lapses et montrons comment estimer l'espace utilisé par différentes classes d'objets dans une pièce.

Pour résumer, les contributions de cette thèse sont les suivantes: (i) nous mettons au point des modèles pour la classification d'image tirant avantage du contexte scénique et des objets environnants et nous proposons une nouvelle base de données pour évaluer leurs performances, (ii) nous développons un nouveau modèle pour améliorer la localisation d'objet grâce à l'observation des acteurs humains interagissant avec une scène et nous le testons sur un nouveau jeu de vidéos comportant de nombreuses interactions entre personnes, objets et scènes, (iii) nous proposons la première méthode pour évaluer les volumes occupés par différentes classes d'objets dans une pièce, ce qui nous permet d'analyser les différentes étapes pour la compréhension automatique de scène d'intérieur et d'en identifier les principales sources d'erreurs.

Abstract

In this thesis, we focus on modeling interactions between people, objects and scenes and show benefits of combining corresponding cues for improving both action classification and scene understanding. In the first part, we seek to exploit the scene and object context to improve action classification in still images. We explore alternative bag-of-features models and propose a method that takes advantage of the scene context. We then propose a new model exploiting the object context for action classification based on pairs of body part and object detectors. We evaluate our methods on our newly collected still image dataset as well as three other datasets for action classification and show performance close to the state of the art.

In the second part of this thesis, we address the reverse problem and aim at using the contextual information provided by people to help object localization and scene understanding. We collect a new dataset of time-lapse videos involving people interacting with indoor scenes. We develop an approach to describe image regions by the distribution of human co-located poses and use this pose-based representation to improve object localization. We further demonstrate that people cues can improve several steps of existing pipelines for indoor scene understanding. Finally, we extend the annotation of our time-lapse dataset to 3D and show how to infer object labels for occupied 3D volumes of a scene.

To summarize, the contributions of this thesis are the following: (i) we design action classification models for still images that take advantage of the scene and object context and we gather a new dataset to evaluate their performance, (ii) we develop a new model to improve object localization thanks to observations of people interacting with an indoor scene and test it on a new dataset centered on person, object and scene interactions, (iii) we propose the first method to evaluate the volumes occupied by different object classes in a room that allow us to analyze the current 3D scene understanding pipeline and identify its main source of errors.

Remerciements

Mes premiers remerciements vont à mes directeurs de thèse Ivan Laptev et Josef Sivic. Vous avez su canaliser mon envie d'explorer de nombreuses pistes différentes pour mener à bien un projet cohérent jusqu'au bout, ça n'a pas dû être facile ! Merci donc pour votre encadrement, pour votre accessibilité, votre implication et votre passion contagieuse. Merci aussi pour avoir su mener à bien ces nombreuses deadlines en gardant la tête froide même 5 minutes ~~avant~~ après l'heure fatidique. Et enfin, merci pour votre bonne humeur en dehors des heures ouvrées, j'ai passé de très bons moments avec vous !

Je remercie également les membres de mon jury. Vittorio Ferrari et Derek Hoeim pour avoir assumé la lourde tâche de rapporteurs, ainsi que Cordelia Schmid, Patrick Perez et Jean Ponce pour avoir accepté de faire partie de mon jury.

Je voudrais aussi remercier tous ceux avec qui j'ai pu travailler. David Fouhey, Abhinav Gupta et Alyosha Efros évidemment: j'ai beaucoup apprécié d'avoir collaboré avec vous et j'ai passé deux excellents mois à CMU en votre compagnie. Mes co-bureaux avec qui j'aurais partagé de nombreuses heures de rigolade et qui ont très largement contribué à faire de ces quatre ans une succession de bons moments: Oliver White, Armand Joulin (sache que mon plus grand regret est de n'avoir pû assister à ton mariage) et *last but not least*, le fameux Nicolas Flammarion ! Le reste de la dream team aussi avec qui j'aurai fait les quatre cents coups: Edouard Grave, Piotr Bojanovski, Guillaume Segin, Rémi Lajugie, Damien Garreau, Fajwel Fogel, Loïc Landrieu, merci pour votre bonne humeur, votre humour noir incisif, vos imitations inimitables ! Spéciale dédicace à Petr Gronat avec qui j'ai partagé un appart et une semaine de Kitesurf. Enfin tout le reste du labo, merci pour les discussions productives que l'on aura pu avoir et votre sympathie: Francis Bach, Rodolphe Jenatton, Augustin Lefèvre, Nicolas Le Roux, Julien Mairal, Guillaume Obozinski, Sylvain Arlot, Florent Couzinié, Warith Harchaoui, Toby Hocking, Matthieu Solnon, Nino Shervashidze, Karteek Alahari, Minsu Cho, Vadim Kantorov, Aymeric Dieuleveut, Sesh Kumar, Anastasia Podosinnikova, Tuan-Hung Vu et Guilhem Chéron.

Enfin, un énorme merci à mes parents grâce à qui j'ai pû arriver là où j'en suis. Merci de vous être inquiété pour ma santé mentale, de m'avoir préparé de bons petits plats (parfois expérimentaux) quand je rentrais le week-end et tout simplement de m'avoir appris à me dépasser. Merci aussi à mon frère et à ma soeur d'avoir fait de ces week-ends des moments de détente pour décompresser un peu. Et puis, évidemment, merci à tous mes amis qui, pour la plupart, auront connu en même temps que moi les joies et galères de la vie de thésard. À tous les membres de la KooCrew,

par ordre de coloc (temporaire ou pas), Dimikoo, Aloïs, Quentin, Alvaro, JR, Fofi, Camikoo: ces années passées avec vous furent du Soignon© pour mes yeux et mes oreilles, et c'est pas fini (comme dirait la pub)! Fofi, chose promise, chose due: merci à toi pour ta bonne humeur indéfectible et contagieuse, ton sourire épanoui et tes gâtés réconfortants, miss you budy ! Un mega merci aussi à tous mes autres potos pour ces week-ends de ouf à droite à gauche: Khém, Toto, Masao, Bromi, Doudou, Romain et Franck. On se sera bien marré ! Et pour terminer, merci à toi, Héloïse, d'avoir hier endossé le rôle de police de la thèse (aidée de Joe le koala), aujourd'hui bien plus, et de m'avoir poussé à terminer d'écrire cette thèse quand la motivation n'y était pas. Tu n'auras peut-être pas assisté à la soutenance mais j'espère que tu pourras assister à tout le reste !

Contents

Chapter 1

Introduction

1.1 Objectives

Studies like [Kourtzi, 2004, Kourtzi and Kanwisher, 2000] in neuroscience have shown that the pose of people is sufficient to perceive the implied motion in still images. This is possible thanks to a part of the brain, the mirror neurons, which are engaged in both the recognition and the realization of actions [Gallese and Goldman, 1998, Urgesi et al., 2006]. It has also been shown that some of those neurons are only activated during a hand-object interaction [Johnson-Frey et al., 2003] but do not respond if only the actor or the object of interest are separately presented to the subject [Gallese et al., 1996]. This seems to indicate a link between the perception of an action and related objects. Other works have actually demonstrated that the recognition of an action is influenced by the scene context [Nelissen et al., 2005, Bach et al., 2005] and that recognizing an object helps recognizing a related action [Bub and Masson, 2006, Chao and Martin, 2000] or vice-versa [Helbig et al., 2006].

Those elements suggest that our computational models should incorporate information

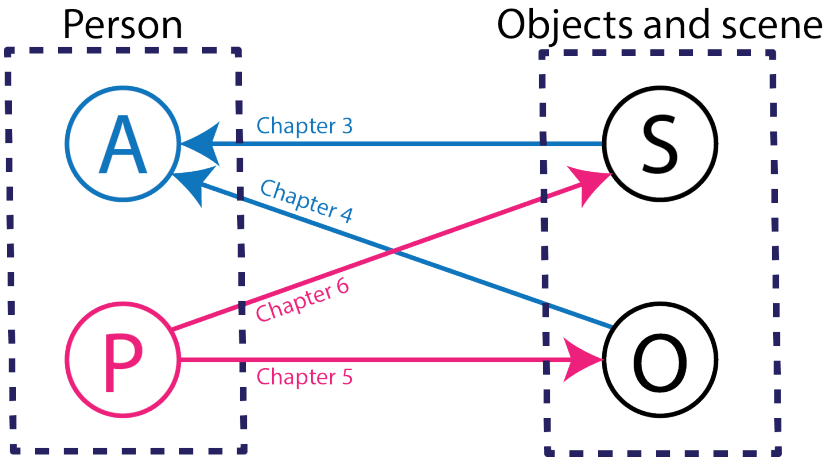


Figure 1.1 – This thesis aims at investigating the interactions between person, comprising action (A) and pose (P), and objects, comprising objects (O) and scenes (S). The arrows point towards the task we seek to improve. The goal of the first part of the thesis (arrows in blue) will be to improve action classification by using scene and object context. The goal of the second part of the thesis (arrows in magenta) is to improve scene understanding and object localization by using the pose of people.

coming from action, pose, objects and the scene. Whereas great progress has been made on action classification, pose estimation, object detection and scene segmentation in the past few years, those research directions have typically been tackled separately. The objective of this thesis is to investigate and model the relations between actions or pose on one side and objects or scene on the other side. More precisely, this thesis is split in two parts, see Figure 1.1 for an illustration:

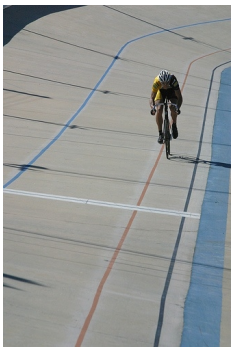
i) In the first part, we want to exploit the action-scene correlation (for example, it is more likely to use a computer indoors than outdoors) and the action-object correlation (e.g. actions like "phoning" or "reading a book" typically involve objects that one can recognize) to improve action classification in still images. The goal of action classification is to identify the action performed by a person among a set of predefined actions. See Figure 1.2 for an example of different action classes for the Willow Action Dataset collected by us in [Delaitre et al., 2010b]. The ground truth annotation usually provides a bounding box around each person of interest present in the image

and the model has to assign one of the possible labels to each box. See Figure 1.3 for three examples of labeled bounding boxes.

ii) J.J. Gibson proposed the notion of affordances [Gibson, 1979]. It describes the way we can perceive the "meaning" or "value" of things in their environment and infer their possible interactions. See Figure 1.4 for an illustration of affordances. In the second part of the thesis, our goal is to exploit the association between a human and the local scene geometry that permits or affords it, in order to infer the underlying 3D structure and semantic object labels of indoor scenes. See Figure 1.5 for an example of using person-object interactions for indoor room semantic labeling.

(a) Using a computer

(b) Riding a horse



(c) Riding a bike

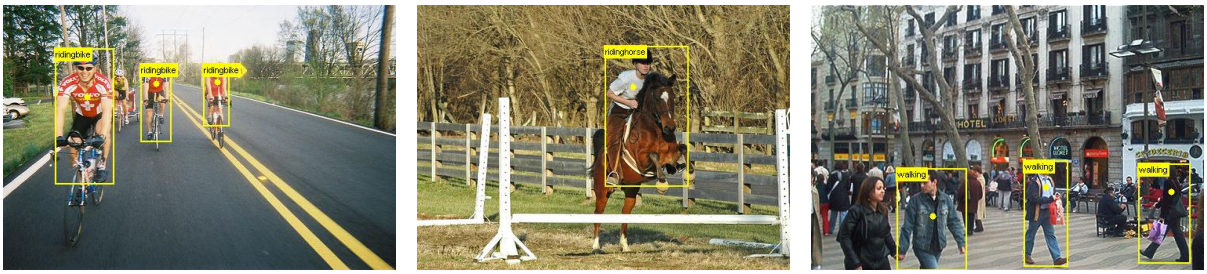
(d) Playing music

(e) Photographing

(f) Running

(g) Walking

Figure 1.2 – The 7 action classes in the Willow Action Dataset [Delaitre et al., 2010b].



(a) Riding a bike (b) Riding a horse (c) Walking

Figure 1.3 – For action classification, the bounding boxes (in yellow) around people are given. The goal is to assign an action label to each box.

(a) (b) (c) (d)

Figure 1.4 – Affordances allow us to infer the function and the use of objects. (a) An example of how affordances help understanding interactions: the round shape of the door handle suggests it can be turned. (b) An example of conflicting affordances: when looking at this object, we feel confused about the way we should use it. (c) and (d): affordances are related to the scale of objects: depending on its dimensions, the same object could be considered as a step (c) or a bench (d).

(a) Interacting with a desk (b) Interacting with a bed (c) Ground truth object labels

Figure 1.5 – People interact with objects in specific poses: (a+b) output of a real pose estimator overlaid over an empty scene without people. Our goal in [Delaitre et al., 2012] is to use the observed person-object interactions to output semantics scene labeling (c).

1.2 Motivation

The number of photos we take has been growing tremendously over the past few years. Estimates report that 380 billion photos were taken in 2012 versus 880 billion in 2014[1], thus an increase of 50% per year. In the same time we share and upload an impressive proportion of this data to the Internet: for example Facebook reported 90 billion image uploads in 2012 (23% of the total number of photos taken that year) and Youtube reached in 2014 the milestone of 100 hours of video uploaded each minute.

People are the generators and consumers of this data and, as such, they appear in a huge proportion of this imagery [Laptev, 2013]. Analyzing so much data has become a challenge by itself (e.g. consider the popularity of the concept "big data") but this data also provides an unprecedented opportunity to design better models to understand how people interact with their environment. Solutions to this task would have a number of important applications:

- **Detecting unusual behaviors:** If it is common to lie on a bed, it might be less frequent to lie on the ground. A finer understanding of how people behave at different places may help to detect endangered persons, e.g. babies or elderly people.

- **Security:** Security in public areas could benefit from good models for recognizing person-object interactions. We would for example be able to detect people running in the street stealing a hand bag or people threatening others with weapons.

- **Person assistance and domotic:** The detection of reaching movements towards an object would have direct applications in robotics or domotic where one could forecast the action of people and automatically start some services. One could for example track all the objects across a whole house to automatically locate an object

[1] Yahoo's flickr 2013 press event.

when someone is looking for it. As objects may be small, it may request a precise understanding of person-object interactions to detect when people pick or drop items under different viewpoints or occlusion.

- **Person interest detection:** Physical stores are interested in a system which would be able to automatically detect people manipulating objects on aisles. This would enable them to track the interest for a product and better design the organization of the aisles to improve the sells.

- **Automatic indexing of large-scale databases:** Improving scene understanding thanks to human interactions would provide useful meta-data for automatic indexing and search of large-scale image archives or datasets. One could for example query Youtube based on the visual content of videos.

- **Build better semi- or unsupervised models:** Great efforts are made to build new datasets with a comprehensive list of classes. The ImageNet 2014 challenge [Russakovsky et al., 2014] for example proposes a dataset with 1.2 million images and 1000 object classes. In parallel, [Dean et al., 2013] gathered a dataset of 32 million images of 100, 000 object classes. Despite the fact that the size of those datasets is impressive, this data still does not cover the rich variety of the real world. To be able to use all the available data, we will need to rely on semi- or unsupervised models where the robust modeling of prior knowledge is crucial. For example, a better model for person-object interactions could help automatic understanding of functional properties of unlabeled objects via affordances.

1.3 Challenges

Automatic understanding of images is a difficult process and the first challenges arise because of the act of shooting a picture itself: two pictures of the same scene can have very different appearance. This could be due to different shooting conditions: viewpoint variations, distance to the object (see how small is the person riding a bike in Figure 1.2(c)), camera's focal length variations, cropping due to the image frame (see the persons interacting with a computer and playing music in Figure 1.2(a) and Figure 1.2(d)) or JPEG compression artifacts. Another source of appearance changes is the scene variation: clothing of people, lighting or background differences, occlusions by nearby "objects" (e.g. in Figure 1.2(b) and 1.2(g): the horse and the man are partially occluded by the water and the board, respectively). Finally, another major challenge is the intra-class variation. See Figure 1.6 for an example of intra-class variation in terms of pose, object, viewpoint and interaction variations.

In addition to these difficulties, modeling the appearance of people in images and video is a very difficult task in itself. In contrast to rigid objects, people can deform in a lot of possible ways, see Figure 1.7. The appearance of a body part is highly dependent of the person's pose and the image of a limb can be drastically foreshortened when projected on the image plane. Limbs can also occlude other body parts. See Figure 1.8 for a visualization of the pose distribution in 3 different datasets. Pose estimation thus remains a challenging problem despite active research on this subject.

There are difficulties specific to action recognition as well. Actions are not well defined: depending on the involved object, the same high-level action can refine into many subcategories. For example, the action "reading" may apply to a book, an e-reader, a map or a sign. In each case, the object and pose involved look very different. See Figure 1.9 for four examples of different ways of sitting. The granularity of an action itself is ambiguous. For

(a) Sitting

(b) Standing

(c) Jumping

(d) Unusual instrument

(e) Unsual viewpoint

(f) Unsual interaction

Figure 1.6 – Intra-class variability for actions may be induced by different factors. (a-c) Changes in pose : people are playing guitar while (a) sitting, (b) standing or (c) jumping. (d) Changes in the object, here a double neck guitar. (e) Changes in viewpoint. (f) Change in the way people execute the action or interact with the involved object.

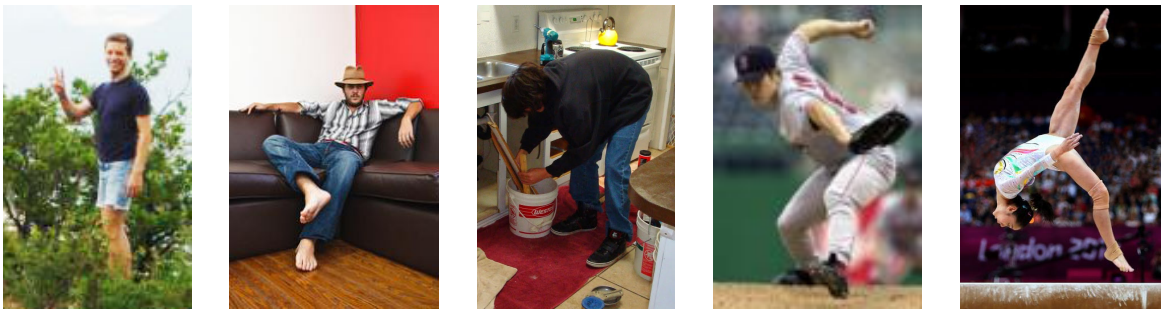
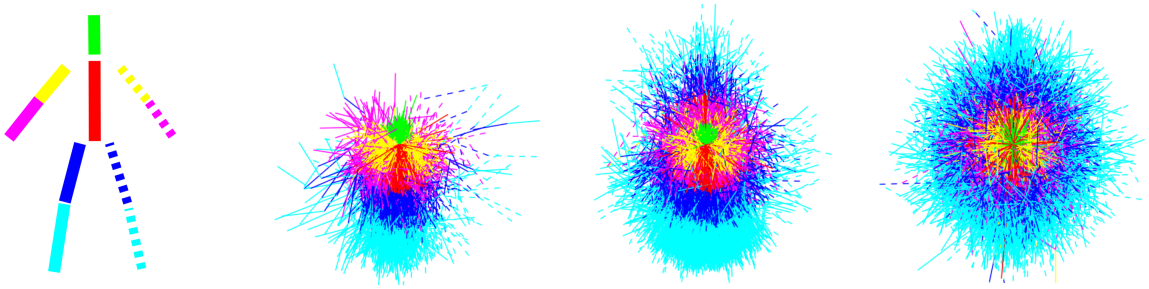


Figure 1.7 – Example of different poses, from our everyday life (left) to less typical poses (right).



(a) (b) (c) (d)

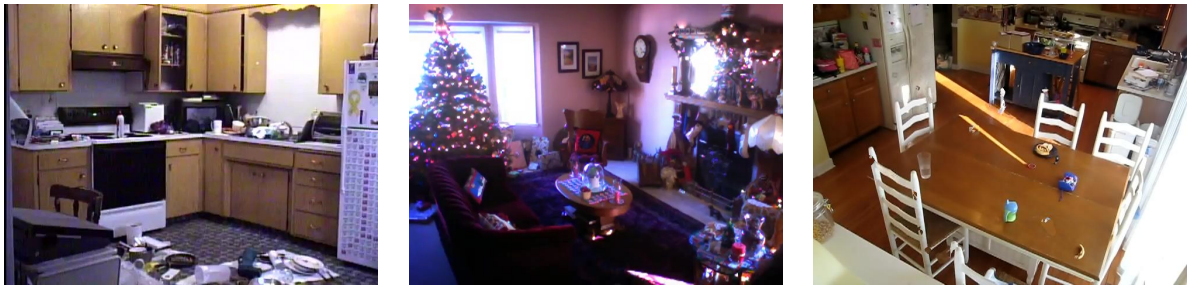
Figure 1.8 – Visualization of the distribution of manually annotated poses for 3 datasets of increasing size and difficulty. The rendering is done by superposing stickmen with the neck at the center of the image: (a) the reference stickman with different colors for each body part. The datasets of: (b) [Ramanan, 2006], (c) [Johnson and Everingham, 2010], (d) [Johnson and Everingham, 2011].

example does it make sense to consider the action of "drinking" or should we decompose it into three successive sub-actions: "raising the glass", "sipping liquid", "putting the glass back"? Due to these ambiguities, ground truth annotations of images might even sometimes be inconsistent.

Finally, the automatic understanding of indoor scenes is again a difficult area. Indoor natural images tend to be very cluttered with lots of severe occlusions. Standard object detectors trained on clean data do not work well in such conditions and have a poor precision. Estimating the room layout of a consumer photograph of an indoor scene (i.e. identifying the different walls, the ground and the ceiling) is hard even under the "Manhattan world"

(a) (b) (c) (d)

Figure 1.9 – Four different poses for action "sitting": (a) sitting on a chair, (b) sitting on a sofa, (c) sitting at a desk, (d) sitting on the ground.



(a) (b) (c)

Figure 1.10 – Example of consumer photographs of indoor scenes. Please note the large amount of clutter on the floor and on the tables in (a), the unusual viewpoint of the sofa in (b) and the severe occlusions of chairs in (c). In all three cases, the intersections between walls and floor are hardly visible.

hypothesis (which assumes that straight lines in a picture are generated by three orthogonal directions) because the lines formed by intersecting walls are often occluded by objects. See Figure 1.10 for examples of cluttered scenes where wall intersections are hardly visible.

1.4 Contributions

The contributions of this thesis can be split in two main parts which are summarized in the following subsections.

1.4.1 Improving action classification in still images from scene and object context

We first demonstrate the efficiency of a locally order-less, spatial pyramid bag-of-features model for action recognition in still images. This model had previously shown to perform extremely well on a range of object, and scene recognition tasks. It represents an image as a distribution of spatially localized "visual" words, obtained by discretizing a set of local SIFT descriptors densely extracted from the image.

We combine the bag-of-features approach with the deformable part-based model detailed in [Felzenszwalb et al., 2009] and demonstrate that improved action recognition performance can be achieved by (i) combining the statistical and part-based representations, and (ii) integrating person-centric description with the background scene context.

We construct a dataset with seven classes of actions in 911 Flickr images representing natural variations of human actions in terms of camera view-point, human pose, clothing, occlusions and scene background. We first evaluate our method on this dataset and show improved performance when taking the surrounding scene into account. We then obtain better results compared to existing methods on the datasets of [Gupta et al., 2009] and [Yao and Fei-Fei, 2010a].

In the second part, we replace the standard quantized local HOG/SIFT features with stronger discriminatively trained body part and object detectors. We introduce new person-object interaction features based on spatial co-occurrences of individual body parts and objects.

We address the combinatorial problem of choosing from a large number of possible interaction pairs by proposing a discriminative selection procedure using a linear support vector machine (SVM) with a sparsity inducing regularizer. Benefits of the proposed model are shown on human action recognition in consumer photographs, outperforming the strong

bag-of-features baseline.

1.4.2 Human pose as a cue for improving scene understanding and object localization

We present an approach which exploits interactions between human actions and the scene geometry to use human pose as a cue for single-view indoor scene understanding. We construct a functional object description with the aim to recognize objects by the way people interact with them. We describe scene objects (sofas, tables, chairs) by associated human poses and object appearance. Our model is learned discriminatively from automatically estimated body poses in many realistic scenes. In particular, we make use of time-lapse videos from YouTube providing a rich source of common human-object interactions and minimizing the effort of manual object annotation.

We show how the models learned from human observations significantly improve object recognition and enable prediction of characteristic human poses in new scenes. Results are shown on a dataset of more than 400,000 frames obtained from 146 time-lapse videos of challenging and realistic indoor scenes.

We then widen our scope and focus on the importance of people for room layout and semantic space occupancy estimation. We extend the ground truth of our time-lapse dataset with 3D locations of camera, objects and walls. We describe a pipeline to estimate the 3D voxels occupied by each object class: camera calibration, room layout selection, 2D object localization and 3D space occupancy. For each step of this pipeline we propose a baseline and an alternative method relying on person detections. Using our 3D ground truth, we show that using people improves over the baseline. We also identify the critical steps of the algorithm and some causes of failure.

1.5 Thesis outline

In Chapter 2, we describe the previous work related to person detection, pose estimation, action classification and human interactions with objects and the scene.

The first part of the thesis focuses on improving action classification from the scene and object context. Chapter 3 details our work on action classification. We first recall the details of the bag-of-features model and the support vector machine classifier. We then analyze the performance of different combinations of parameters, including: the feature choice, the vocabulary size, the spatial pyramid and the choice of the kernel. We demonstrate the benefits of incorporating scene context information and combine the orderless bag-of-features model with the deformable part model of [Felzenszwalb et al., 2009]. We finally compare the proposed approach on a set of 4 challenging datasets. This chapter is based on a publication at BMVC 2010 [Delaitre et al., 2010a].

Chapter 4 extends the action classification method described in Chapter 3 and introduces new features based on the discovery of pairs of correlated person/object detectors. To avoid the combinatorial explosion of pair candidates, we discriminatively select the pairs of person and object detectors by training a classifier with a sparsity inducing norm. We first detail the model and then show improved results over the strong bag-of-features baseline. Results of this chapter were published at NIPS 2011 [Delaitre et al., 2011].

The second part of the thesis focuses on improving scene understanding from pose estimation of actors interacting with the scene. In Chapter 5, we describe a semi-supervised model for indoor scene understanding. More precisely, we first gather a set of time-lapse videos, i.e. videos sparsely sampled in time with fixed background, showing numerous interactions between people and the annotated environment. We gather poses using the model of [Yang and Ramanan, 2011]. We use this data to automatically learn how people interact with different categories of objects such as “sofa”, “chair” or “table”. We first

describe the model and then show significant improvements over the baseline using only visual cues. This work was published in ECCV 2012 [Delaitre et al., 2012].

We extend both Chapter 5 and our second publication [Fouhey et al., 2012] in ECCV 2012 (not included in this thesis) in Chapter 6 to investigate the role that people can play in scene layout and occupied space estimation. We add 3D ground truth annotations to our time-lapse dataset and show that using people can help at each steps of the algorithm. We also identify the critical steps responsible for the biggest performance drops and discuss the possible causes of failure.

We conclude this thesis in Chapter 7 by a summary of the contributions and outline the perspectives for future work.

1.6 Publications

- Delaitre, V., Laptev, I., and Sivic, J. (2010a). Recognizing human actions in still images: a study of bag-of-features and part-based representations. In Proc. BMVC. updated version, available at http://www.di.ens.fr/willow/research/stillactions/

- Delaitre, V., Sivic, J., and Laptev, I. (2011). Learning person-object interactions for action recognition in still images. In NIPS

- Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Efros, A. A., and Gupta, A. (2012). Scene semantics from long-term observation of people. In ECCV

- Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A. A., Laptev, I., and Sivic, J. (2012). People watching: Human actions as a cue for single-view geometry. In ECCV (not included in this thesis)

Chapter 2

Literature review

There are numerous ways one could exploit interactions between people, scenes and objects. One can use the fact that the presence of people is correlated with the presence of objects. For example a person is more likely to walk on the sidewalk than on the road. It can also happen that the pose of the person itself is characteristic for an object, e.g. when people casually sit on a sofa, or simply that the action performed by the person implies the use of a specific object, e.g. to play tennis one needs a racket. For those reasons we review the literature related to person detection in Section 2.1, to pose estimation in Section 2.2, to action classification in Section 2.3 and to scene understanding in Section 2.4 before detailing the literature on interactions between people, objects and scenes in Section 2.5.

2.1 Person detection

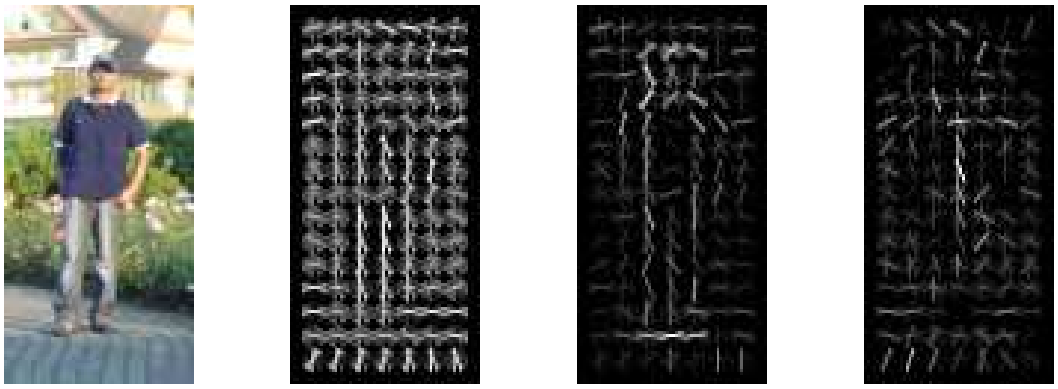
Person detection aims at localizing the position of people in images or videos. Besides the viewpoints and lighting variations, the models also have to accommodate the highly deformable nature of the human body which makes this task very challenging, especially

in scenes where multiple persons stand close to each others.

Early works on video were exploiting the movement of people to segment and detect them [Saptharishi et al., 2000, Jabri et al., 2000, Krahnstoever and Mendonca, 2005, Rother et al., 2007]. By building a model of the background using median images, auto-regressive filters or Gaussian models, these methods are able to identify and group foreground pixels into a person mask, thus performing person detection via segmentation. Although these simple methods are suited for real-time detection, they suffer from the following two major limitations: the shadow casted by pedestrians will be associated with the person and static people will be merged with the background.

As the shape of a standing pedestrian is quite characteristic, other approaches explored the possibility to match image edges with person templates. Those templates are either manually annotated clean outlines of people's silhouettes [Gavrila, 2000] or 2D-filters on contours learned from data by a perceptron [Felzenszwalb, 2001]. Due to the unreliable nature of edge extraction in presence of clutter and the high variability of pose contours, those methods are only suited to controlled environments with a uniform background, no occlusions and low pose variations.

A class of methods was initiated in parallel by [Papageorgiou and Poggio, 2000] and [Oren et al., 1997]. They propose a sliding window detector which extracts features based on Haar wavelet transforms from a rectangular image patch and use a SVM classifier to decide if the current bounding box contains a person or not. Thanks to the use of more discriminative descriptors and machine learning, those methods are more robust to clutter. The work of [Dalal and Triggs, 2005] introduces a similar detection algorithm based on Histograms of Oriented Gradients (HOG) descriptors instead of wavelets. Those descriptors are computed from the discretization of the intensity gradient into several orientation bins and normalized by the local image contrast. See Figure 2.1 for a visualization of the HOG descriptors and the SVM-based person detector weights. This descriptor remains in use



(a) (b) (c) (d)

Figure 2.1 – Visualization of the HOG descriptors and SVM-based person detector weights from [Dalal and Triggs, 2005]: (a) Original image, (b) HOG descriptors visualization (dominant edge directions), (c,d) HOG descriptors weighted respectively by the positive and negative weights of the SVM classifier.

by the state-of-the-art methods such as [Yang and Ramanan, 2013].

The drawback of those methods resides in the fact that they rely on extracting features from a rigid image patch covering the full body and thus are not able to successfully capture the body pose variations. An extension was thus proposed in [Mohan et al., 2001] where the human body is represented by four components, each being associated with a classifier: upper body, left arm, right arm and lower body. Those components are allowed to move within the person bounding box but their extent is defined manually. This method introduces less rigid person detectors but remains ad-hoc and suited only for standing front-facing people.

The previously discussed methods could actually apply to any object. They do not formulate the detection problem as a pose estimation problem and neglect the kinematic properties of the body parts. The method proposed by [Ioffe and Forsyth, 2001] tries to take advantage of body structure and fit oriented rectangles for each body limb using kinematic constraints. Fitting rectangles in an image is very noisy but this is a first step towards combining detection and pose estimation. Another method based on body

part classifiers [Mikolajczyk et al., 2004] estimates an assembly of parts from coherent part detection configurations.

The two previous papers select the part assembly in a greedy manner, using a multi-stage pipeline and discarding hypothesis which are not plausible enough according to spatial criteria. The pictorial structure model introduced by [Fischler and Elschlager, 1973] overcomes the difficulties related to those numerous thresholding steps. This model represents the image of an object as a set of parts spatially organized in a deformable configuration and allows to compute a likelihood for any part configuration. It was made computationally tractable by [Felzenszwalb and Huttenlocher, 2005] thanks to the generalized distance transform [Felzenszwalb and Huttenlocher, 2004] used to compute the binary potential representing the relative displacement of parts. In the case of person detection, each part captures the local appearance of some body region and is able to slightly move away from its anchor position, which is particularly suited for body limbs. In [Felzenszwalb and Huttenlocher, 2005], the problem is formalized using a generative approach and the parameters are learned by maximum likelihood. In the later works [Felzenszwalb et al., 2008b] and [Felzenszwalb et al., 2009] a discriminative formulation is introduced which allows to make use of negative examples to learn a more robust model. In addition, training images are separated into clusters of similar aspect and a specific pictorial model, called component, is learned for each cluster to account for variations in viewpoint and pose.

Despite the use of multiple components, usually not more than a dozen, the deformable part model proposed by Felzenszwalb cannot capture the very rich variety of poses. [Bourdev and Malik, 2009] tackles this issues by collecting many images of people, grouping the ones having a *locally* consistent pose together and training a classifier for each of those local "poselets", see Figure 2.3. It would, for example, group people with crossed arms in a cluster, no matter what the lower body configuration looks like.

(a) (b) (c)

Figure 2.2 – The person detector from [Felzenszwalb et al., 2009]: (a) The root filter, similar to the person detector of [Dalal and Triggs, 2005] (see Figure 2.1), (b) weights of the five deformable parts of the person model placed at their anchor position, (c) deformation cost payed when placing parts away from their anchor position: the darker the closer to zero.

[Bourdev et al., 2010] then proposed a method to aggregate the output of those poselets into a person detection algorithm. This method performs very well thanks to the use of many detectors specific for a given pose but it is computationally demanding.

All the methods we detailed so far do not model occlusions. This can be a limitation as people often appear in complex scenes with severe occlusions. They might be for example partly hidden by other people, e.g. in street-like environments, or by furniture, e.g. in indoor scenes. The method introduced by [Girshick et al., 2011] builds upon the deformable part model and includes a specific "occluder" part. Based on a manually defined grammar, the model parses the scene to find the most likely configuration of parts.

Instead of having a part based model with multiple components describing the full-body appearance, it is also possible to have multiple components for each part. The pose estimation model of [Yang and Ramanan, 2011] follows this approach and proposes a tree structured model with a part for each body joint and limb, each having several different components with a specific spatial configuration: for example a "folded elbow component" does not expect the hand to be at the same place as a "straight elbow component".

Figure 2.3 – Example of positive images used to train poselet detectors. From top to bottom: frontal face, right arm crossing torso, pedestrian, right profile and legs. Note how images are correlated in both appearance and body configuration.

The extension of [Desai and Ramanan, 2012] also handles part occlusions. According to [Yang and Ramanan, 2013], this model performs slightly better than the deformable part model for person detection but the real advantage is that it provides an estimation of the body pose at the same time.

Person detection was used in our work [Fouhey et al., 2012] (not included in this thesis) to detect walkable, sittable and reachable surfaces in indoor environments. We chose the detector of [Felzenszwalb et al., 2009] for its high precision and re-trained it on three types of images depicting standing, sitting and reaching people. We then cast votes for each detection to evaluate the extent of the corresponding surfaces.

2.2 Human pose analysis

In this section, we review the literature related to body pose estimation, i.e. to the estimation of the body joint positions. In addition to the difficulties encountered for person detection, models are prone to errors due to the body self-occlusions and the correlation of appearance between the left and right limbs. Technically, the task is also difficult because of the high number of parameters needed to represent a body pose and the fact that their underlying probabilistic distribution are non-Gaussian and multi-modal [Deutscher et al., 1999]. Early work on pose estimation followed the same evolution as for person detection and started to focus on videos using features based on background subtraction and motion cues before turning to still images when more robust features, e.g. HOGs [Dalal and Triggs, 2005], appeared.

When it comes to estimate the position of the limbs of a person in an image, a solution can be to model the body as a 3D assemblage of cylinders. Those cylinders are linked by joints with a certain number of possible rotations and can move with respect to each other. A realistic model of the human body usually has at least 20 degrees

of freedom and the goal is to estimate the value of the corresponding limb rotations and translations depicted in an image. The transformations between the measurements (pixel positions in image plane) and the variables (rotation amount around a joint axis) are highly non linear and are not related by a 1-1 mapping as multiple 3D configurations can project to very similar pixel measurements. The problem is thus ill-posed but can be addressed by imposing certain constraints on the solution. Early approaches like [Deutscher et al., 1999, Deutscher et al., 2000, Sidenbladh et al., 2000] adopted a generative approach and associated a random variable to each degree of freedom. They could however not use traditional statistical modeling methods such as conditional random fields due to the size of the search space and non-Gaussian aspect of conditionals. To overcome those issues, they used a sampling technique known as particle filtering [Gordon et al., 1993, Isard and Blake, 1998] to represent the distribution of the joint angles in each video frame.

Another class of methods directly learns a mapping from the 2D projected image joint positions to the 3D model parameters. The intuition behind those approaches is that though living in a high dimensional space, the physically plausible human body configurations are very sparse and could be embedded in a much smaller space. This is the case of [Elgammal and Lee, 2004] who use non-linear dimensionality reduction such as Isomap [Tenenbaum et al., 2000] or Local linear Embeddings [Roweis and Saul, 2000] to learn a mapping from the image to the 3D pose. Such dimensionality reduction methods are however prone to overfitting, require lots of training data and do not directly provide a mapping between the embedding space and the pose space. A more robust method based on Gaussian Process Latent Variable Models (GPLVM) was introduced by [Lawrence, 2003] and used in [Grochow et al., 2004, Urtasun et al., 2005] for 3D pose tracking. Those methods compute a smooth prior in the embedding space allowing them to competitively reduce the reconstruction error. They were further extended to include temporal smoothing in

the embedding space [Urtasun et al., 2006], local preservation of the structure of the pose space [Urtasun et al., 2007, Kanaujia et al., 2007, Hou et al., 2007] and local preservation of the structure of the feature space [Gupta et al., 2008a].

The previously cited methods allow to successfully track limbs positions in videos but do not solve pose estimation in itself as the models based on dynamics need to be manually initialized. [Agarwal and Triggs, 2004] proposes a solution based on learning a non-linear regression from the silhouette to the body pose. Using background subtraction this method proved promising in videos with some cluttered background but cannot disambiguate the left/right symmetries as it relies only on the body contours. Other approaches like [Sigal et al., 2004, Gupta et al., 2008b] use the constraints generated by multiple views of the same scene to automatically estimate the pose. This is however a very particular setup as most of the available data is single-view. A more general method was developed by [Ramanan et al., 2005]. It uses a person detector to detect a person, learns instance specific classifiers on RGB pixels for each body part and then tracks the person's pose in the full video sequence using the pictorial structure framework detailed in [Felzenszwalb and Huttenlocher, 2000]. [Ramanan, 2006] further extended this work by replacing the person detector with an initial pose estimation computed from weak edge-based body part classifiers and iteratively learns the appearance model of individual parts. This idea was adapted to video by [Ferrari et al., 2008a].

The deformable part model of [Felzenszwalb et al., 2008b] was the basis of a new class of discriminative methods for pose estimation in images. While [Andriluka et al., 2009], [Sapp et al., 2010] and [Andriluka et al., 2010] focused on speeding up the inference and training more robust part detectors, [Johnson and Everingham, 2010] brought the interesting new idea that the appearance of parts is multi-modal which translated into new models with multiple classifiers or "components" per part. For example, the method of [Johnson and Everingham, 2011] splits the training poses into clusters having similar pose

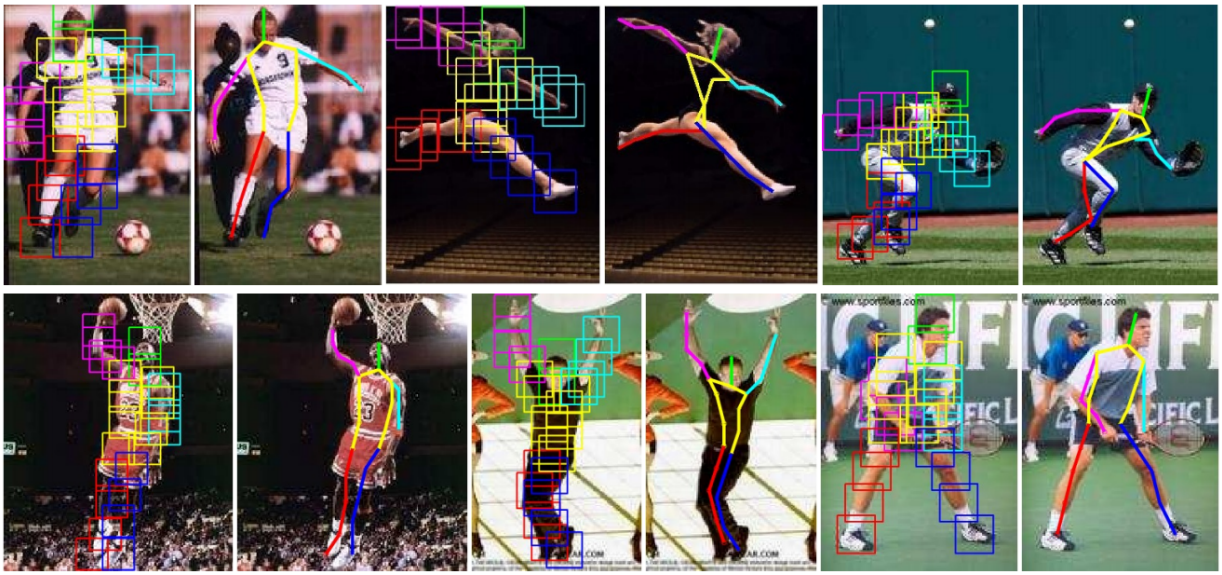


Figure 2.4 – Some pose estimation examples for the model of [Yang and Ramanan, 2011].

configurations and learns multi-modal part classifiers within each pose cluster. On the contrary, the method of [Yang and Ramanan, 2011] does not have any global pose cluster but the position prior of each part is conditioned on the component of the parent part. The expected position of parts is thus directly related with their appearance which makes this model very powerful. See Figure 2.4 for some detection examples.

The understanding of importance of having a position prior depending on the appearance of mid-level patches allowed new directions of research for pose estimation. Some works improve regressions from the part appearance to children part positions: [Dantone et al., 2013] uses random forests to estimate the joint position prior from the appearance features and the prior of other joints. [Hara and Chellappa, 2013] learn non-linear regressors from image features to the displacement of each child part thanks to regression trees. Another possibility is to use the "poselets" of [Bourdev and Malik, 2009] to generate the joint position prior from poselet activations: [Gkioxari et al., 2013] applies this idea to the arm and [Pishchulin et al., 2013a, Pishchulin et al., 2013b] to the whole body configuration. Although successful, those works optimize the unary terms of their

model separately from each other. On the contrary, [Yang and Ramanan, 2011] managed to formulated his optimization objective as a structured cost over all the joints which was also part of the success of this model. The model of [Wang and Li, 2013] exploits the idea of jointly training the model parameters for all body parts and introduces combined-parts which model appearance and joint positions at a bigger scale. A similar approach was developed by [Sapp and Taskar, 2013] who jointly train high and mid-level body part detectors in a structured manner.

Those last approaches are however outperformed by the methods which take advantage of the recent advances in deep learning. [Toshev and Szegedy, 2014] use a cascade of deep convolutional networks to learn a regressor from the image patches to the joint positions. An even more efficient method introduced in [Tompson et al., 2014] jointly learns the parameters of a deep network and a Markov random field to estimate a likelihood map of the body joints. In addition of being state-of-the-art in 2D pose estimation, the inference of those two last methods is one or two order of magnitude faster than all the previously cited methods.

Pose estimation was at the center of our work [Delaitre et al., 2012] where we used people as a prior for object location in indoor rooms. In detail, we improved classification of super-pixels into different object classes by designing a feature inspired by a bag of visual words. The feature relies on discretization of the human pose into clusters and encodes the distribution of poses around super-pixels. We also produce a visualization of the learned classifier weights and observe the expected spatial distribution of object locations around each of the pose clusters.

2.3 Human action recognition

The goal of action classification is to identify the action performed by a human actor in a video or a still image. The set of possible actions is pre-determined and the typical classes of interest usually involve daily actions such as *walking, running, eating, drinking, phoning, taking a picture, reading a book, playing an instrument.* It is generally assumed that depicted people occupy a large fraction of the image. If it is not the case one can use a person detector to focus on the area of interest.

Due to the association between action, motion and pose, early work in action classification focused on video. It was mostly based on extracting the statistics of the evolution of the silhouette of people over time [Bobick and Davis, 2001]. Methods based only on the silhouette cannot however precisely describe the action of a person. Driven by the success of the local invariant features in object recognition [Lowe, 1999, Hall et al., 2000], the works of [Laptev and Lindeberg, 2003] and [Dollár et al., 2005] introduced space-time interest points for video. Combined with a clustering algorithm like K-means, it allows to represent a video by a set of quantized local space-time features called "visual words" in analogy to words in text. For example, [Schuldt et al., 2004] developed a method based on quantized features to classify actions in videos thanks to the bag of visual words representation and a SVM classifier. Others [Wong et al., 2007] adapted topic models to include spatial information of visual words whereas the method of [Niebles et al., 2008] uses them to associate a latent topic to each class and do action classification in an unsupervised manner. A different kind of approach based on neural networks was developed by [Jhuang et al., 2007]. They explicitly include two layers to deal with motion and temporal variability of actions.

Those last methods were tested on the KTH dataset [Schuldt et al., 2004] and the Weizman dataset [Gorelick et al., 2007] which both contain sequences with simple backgrounds

and static cameras and respectively have 6 and 9 action classes. Those datasets do not show actions performed in real-world environments so [Laptev and Pérez, 2007] proposed the 95 minutes long movie "Coffee and Cigarettes" with two actions: "drinking" and "smoking". Although already challenging because of the scene clutter and because the two classes are close to each other, this dataset is limited only to 2 classes. [Marszalek et al., 2009] thus mined common actions from 69 Hollywood movies and created the Hollywood2 dataset with twelve action classes: "Answer Phone", "Drive Car", "Eat", "Fight Person", "Get Out Car", "Hand Shake", "Hug Person", "Kiss", "Run", "Sit Down", "Sit up", "Stand up". It remains one of the most challenging datasets for action classification in videos, along with HMDB51 [Kuehne et al., 2011] and UCF101 [Soomro et al., 2012] with 51 and 101 action classes, respectively. State-of-the-art methods for action classification in videos include the method of [Wang and Schmid, 2013] based on the dense extraction and encoding of trajectories and the method of [Simonyan and Zisserman, 2014] based on deep neural networks trained on both the spatial and temporal video domain.

The previously cited actions show clear movement but some actions like *taking a picture*, *reading a book*, *using a computer* or *playing instrument* can be static or involve only subtle movements. There is thus really a need to model actions in static images and we focus on still images in this thesis.

Following the development of action recognition in video, the first works in still images were based on extracting visual descriptors from the silhouette of people [Ikizler et al., 2008] or matching the contours of two silhouettes [Wang et al., 2006b]. Those approaches can however capture only a very coarse representation of actions and are sensitive to occlusions and change of viewpoint. On the contrary, methods relying on appearance based descriptors are more keen to represent specific configurations of overlapping body parts and proved more robust. [Ikizler et al., 2009] therefore extended the person detector introduced by [Dalal and Triggs, 2005] to learn a representation of actions in still images in

the form of a rigid HOG-template.

This method however neglects the deformable aspect of the human body and we tried to address this issue in [Delaitre et al., 2010a] by combining the deformable part model of [Felzenszwalb et al., 2009] with the locally order-less representation of spatial pyramids [Lazebnik et al., 2006] to use both structured and statistical cues. Another possibility is to use poselets to detect the local gestures which characterize an action [Yang et al., 2010, Maji et al., 2011]. The work of [Yao et al., 2011a] goes further by using the "activation vector" of a set of pre-trained "poselets", visual attributes and object classifiers to represent an image. Their method directly learns a classifier to recognize the action using the output of these classifiers and also computes a sparse representation of the image by decomposing the activation vector using a sparse "action basis" via dictionary learning. This action basis representation copes well with the false detections of objects or attributes and shows improved performance. See Figure 2.5 for an illustration of the learned action bases.

The above discussed works rely on standard classification techniques and do not actually *model* actions. They do not exploit the very specific nature of the task: most actions involve the interaction of a body part with another body part, an object or the surrounding scene. Some works have pushed in this direction. For example, [Yao and Fei-Fei, 2010a] designed an algorithm to discover Grouplets, i.e. groups of features in a given configuration that characterize the action. In the same manner, [Yao et al., 2011b] proposed a method to learn pairs of correlated patches for action recognition.

We conclude this section with the recent work of [Oquab et al., 2014] who apply deep convolutional networks to action recognition and obtain state-of-the-art results. We now review the literature related to scene understanding before reviewing the work about person, object and scene interactions in Section 2.5.

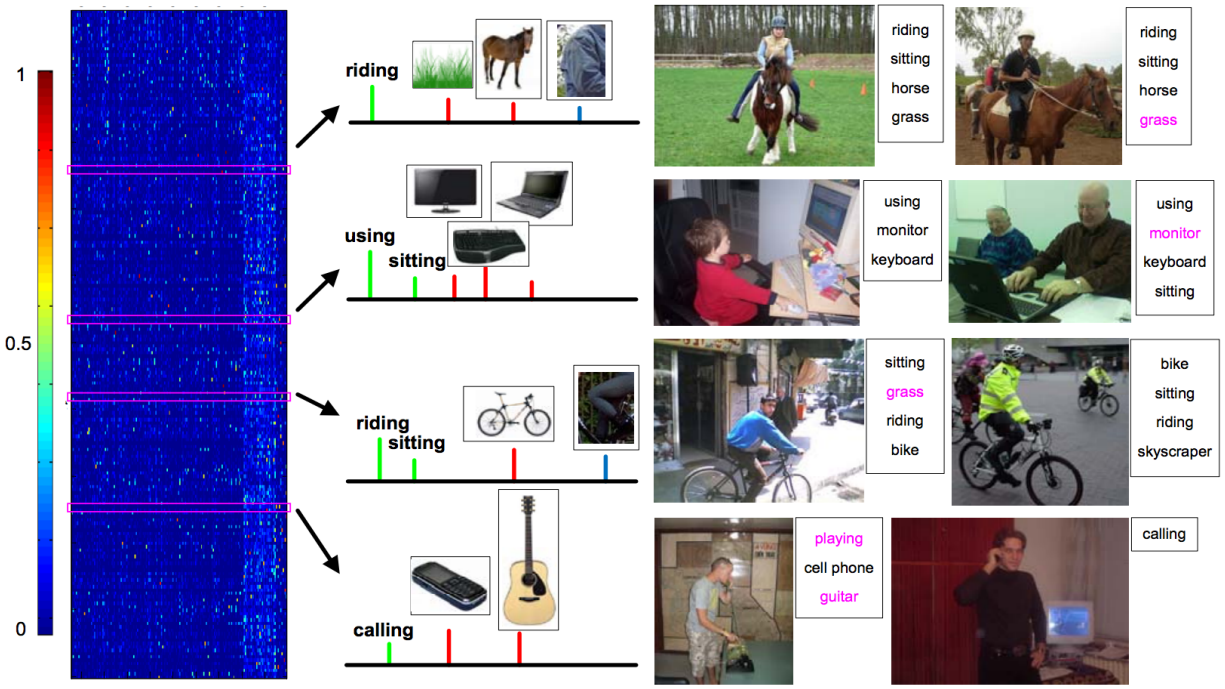


Figure 2.5 – Visualization of the action bases of [Yao et al., 2011a]. The left-most part represents the matrix of action bases (one per row) with color going from blue for a null coefficient to red for 1. The middle part shows the attributes (green bar), objects (red bar) and poselets (blue bar) associated with selected action bases. Note the sparsity of the learned basis. The right-most part shows images where the corresponding four bases have a high contribution as well as a list of inferred labels (attributes and objects). Incorrect labels are in magenta.

2.4 Scene understanding

The notion of scene understanding embraces a set of more specific task, from scene classification to room layout estimation. These different tasks however share a common difficulty: the aspect of scenes varies a lot. Scenes can be highly structured like indoor rooms where lines in the image tend to be generated by three orthogonal directions of the space, or less structured like a landscape picture. Conversely, scenes may also have a global common aspect but may subtly differ by the arrangement of their layout (e.g. a corridor versus an indoor room) or by the presence of specific objects (e.g. a living room versus a dining

room). In this section we review some specific tasks that belong to the more general area of scene understanding.

One of the tasks related to scene understanding is scene classification. The goal is to classify scenes into different categories. [Vogel and Schiele, 2004] proposed a two-stage classifier to discriminate between different natural scene types. The first stage classifies image regions into semantic classes like "sky", "water", "foliage" or "flower". The second stage classifier then outputs the scene type based on the position and fraction of image covered by each semantic class. This method however requires to annotate the different classes involved to train the first stage classifier. To avoid this annotation step, [Fei-Fei and Perona, 2005] describe a Bayesian model for classifying an image into 13 scene categories based on the output of simple image filters. [Quattoni and Torralba, 2009] developed a better method relying on a max-margin classifier and showed that the GIST scene descriptor introduced by [Oliva and Torralba, 2001] obtained similar performance to a bag-of-feature models. As mentionned earlier, some scenes may be characterized by the presence of specific object, e.g. a library can be recognized by the presence of lots of books. This cannot be captured by the coarse GIST descriptor or by the orderless bag of visual words model. [Pandey and Lazebnik, 2011] thus adapted the DPM of [Felzenszwalb et al., 2009] for scene classification. This representation however suffers from being too structured: two scenes of the same class may not share the same objects or spatial arrangement. People have thus looked at intermediate representations. [Li et al., 2010] propose a representation based on the output of a bank of several pre-trained object detectors. More recently, [Doersch et al., 2013] proposed a method to discover and train specific mid-level visual element which are both representative, i.e. frequently occurring within the scenes, and discriminative, i.e. able to differentiate between the scenes.

Another task related with scene understanding is scene segmentation. The goal is to partition the image into a set of regions which correspond to object classes in the

scene, see Figure 2.6. This was first addressed by [Ohta et al., 1978] which followed a top-down *ad hoc* procedure based on manually defined production rules to group and label images patches into regions according to their context. Later and more general approaches adopted those notions of over-segmentation and context by using graphical models to group neighboring pixels [Shotton et al., 2006], super-pixels [Kohli and Torr, 2009] or densely extracted patches [Fei-Fei and Li, 2010] into semantic images regions. Others [Tighe and Lazebnik, 2010] investigated non-parametric methods and label super-pixels according to the most similar super-pixels in a set of visually similar annotated images. Those methods are however limited by the fact they cannot reason about object instances. [Silberman et al., 2014] thus represented an image segmentation as a cut of a hierarchical segmentation tree and proposed to train a structured model on a dataset annotated at the object instance level. In this approach the actual segmentation of the image is as important as the pixel-wise labeling of the image: e.g. two neighboring region representing cars are not merged into one big "car" region in the ground truth. The previous methods do not however use the structure of the scenes in images. One could for example use the fact that most scenes contain horizontal planes, e.g. the ground or the top of a table, and vertical planes, e.g. buildings or walls. This has been recently investigated by [Khan et al., 2014] who apply scene segmentation specifically to indoor scenes by enforcing consistency of the labeling for pixels which belong to the same detected planar regions.

This idea of understanding the geometric structure of a scene has been addressed by many people. [Hoiem et al., 2005] developed the Geometric Context descriptor which computes the probability that a pixel belongs to the ground, the sky, a vertical surface (either front, left or right facing) or an object (either solid or porous). This was for example used by [Gupta et al., 2010] to describe an image by a set of blocks whose depth ordering is consistent with the image. Others [Barinova et al., 2010] tried to estimate the horizon line and the vertical vanishing point. This could be used to help removing false positives for object

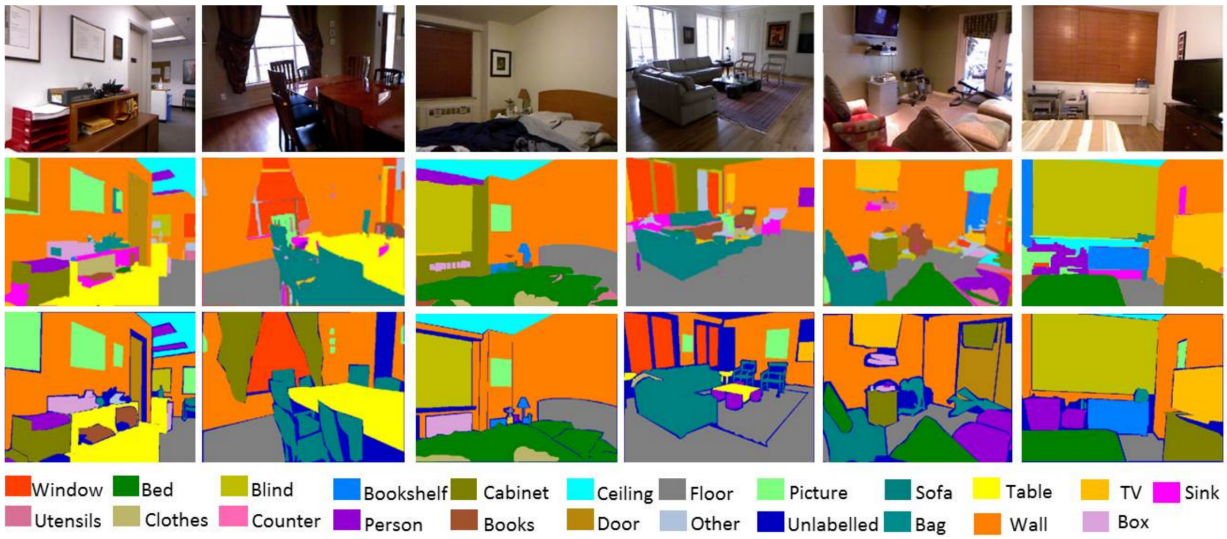


Figure 2.6 – Results of scene segmentation from [Khan et al., 2014]: (top row) input image, (middle row) ground truth labeling, (bottom row) segmentation results.

detection. [Bao et al., 2011] for example addressed this problem by developing a method to estimate a plane supporting multiple object detections to remove false positives. Following the same goal, [Payet and Todorovic, 2011] estimated the coarse 3D shape of the scene from its texture and improved object detection by keeping only detections which agree with the scene.

A finer automatic understanding can be achieved by focusing on a specific type of scene. For example, [Geiger et al., 2011] focused on street scenes and propose a model to recover the street width, the number of intersections at a crossing, the angle between intersections and the image segmentation into “sky”, “background” and “street regions”. Such specific work on outdoor scenes remains however seldom and people gave more attention to indoor scenes as the environment is more controlled. Under the Manhattan world assumption stating that most of the lines in an image are generated by 3 orthogonal directions, it is indeed possible to recover the 3 corresponding vanishing points. Using this idea, [Lee et al., 2009] proposed an image feature called Orientation Map which is used to find surfaces perpendicular to each of the 3 vanishing directions of the scene. They rely on

it to estimate the layout (i.e. the wall positions) of the room by iteratively adding corners and selecting the one hypothesis whose walls best agree with the Orientation Map, see Figure 2.7(b). This orientation map has been recently improved by the methods developped in [Fouhey et al., 2013] and [Fouhey et al., 2014] by transfering the surface orientation from RGBD data and imposing contraints on the relative orientation of different image regions. [Hedau et al., 2009] also address the problem of layout estimation but assume that room are parallelepipeds. They generate many room hypothesis from the detected vanishing points and train a structured predictor to select the best one. This work was further extended by [Wang et al., 2010] by adding latent variables to account for the presence of clutter and enrich the features. Our work [Fouhey et al., 2012] also builds on [Hedau et al., 2009] by including a penalty term based on person detections. Its goal is to discard room hypothesis such that detected people do not fit on the floor. [Chao et al., 2013] also used people for the same goal by using a prior on the 3D person height for standing and sitting people and discarding inconsistent room hypothesis.

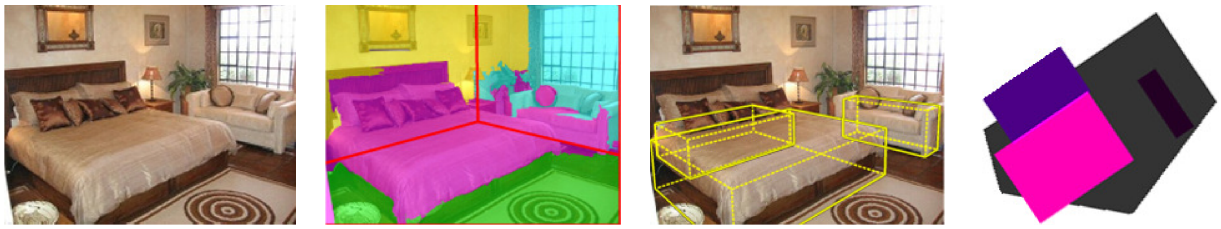
Some methods also try to detect the location of objects in addition to the room layout. [Lee et al., 2010] and [Schwing et al., 2012] generate object hypothesis thanks to the Orientation Map of [Lee et al., 2009] and add a penalty for layouts where objects are intersecting or outside the room, see Figure 2.7(c). [Hedau et al., 2010] adopt a similar strategy but generate objects randomly and take into account the distance between walls and objects. This method was further extended in [Hedau et al., 2012] to take the room context into account. [Del Pero et al., 2011] and [Del Pero et al., 2012] adopt a different approach. They model the room as a collection of objects represented by boxes and define a likelihood relating image features and a room hypothesis. They sample different rooms using the Markov Chain Monte Carlo algorithm a return the most likely. Another method [Zhao and Zhu, 2011] uses a stochastic grammar which decomposes a scene into an arrangement of boxes (stacked or aligned boxes) and faces (nested or aligned faces).

This description of indoor scenes allows to parse (in the sense of formal language theory) the picture of a room and extract the room layout and the boxes it contains.

The previously cited methods are not able to identify the category of the detected objects and treat them as clutter whereas the method proposed by [Satkin et al., 2012] can recover the labels of objects in the room. It relies on a database of room examples annotated in 3D. It learns a ranking between an input image and the 3D annotation in the database and transfers the scene annotation from the top scoring example. [Choi et al., 2013] adopt a different approach. They compute a set of 3D geometric phrases which represent the set of common arrangements of 3D objects in the training images. These are obtained by generating all the possible interactions between objects and clustering the candidates which appear in multiple images. They then optimize an energy with object-scene and object-object compatibility (based on the 3D geometric phrases) by sampling and show improvements over layout estimation and object detection. As this method relies on detections obtained by a deformable part model [Felzenszwalb et al., 2009], it dramatically fails in highly cluttered rooms or in the presence of people occluding the objects. The previously cited works investigated the interactions between objects and scenes. In the next section, we review interactions between persons, objects and scenes.

2.5 Person-object-scene interactions

In this section, we review the related work on interactions between people and their environment. People, actions, objects and scenes are correlated and each of these clues can be used as a context for the others. For example, the action “riding a bike” is very likely to happen outdoors and one should detect a person on a bike. Some information can thus by gained by modeling such relationships and this usually results in a gain in recognition performance. In the following we first explore the interactions between people and manip-



(a) (b) (c) (d)

Figure 2.7 – Room layout estimation from [Hedau et al., 2012]. (a) Input image. (b) During the first step of the algorithm, super-pixels are assigned to either clutter (in magenta), ground (green), left wall (not present here), middle wall (in yellow), right wall (in cyan) or ceiling (not present here). The optimal wall boundaries (in red) are then computed. (c) Candidate object 3D bounding boxes. (d) Top view of the 3D boxes. [Satkin et al., 2012] propose to use the average precision of the free space (in gray) as a performance measure.

ulable objects (Subsection 2.5.1) and then review work on interactions between people and scenes (Subsection 2.5.2).

2.5.1 People and manipulable objects

Manipulable objects are usually small and thus hard to detect and classify. But by detecting the position and motion of a person's hand one can estimate a prior on the position of the object. Similarly, by recognizing the action of a person, one can also obtain a prior on the object category. In the following we review related work exploiting these kinds of interactions.

The first type of approach uses interactions between hands and the environment as a cue to detect objects. For example, the method proposed in [Moore et al., 1999] uses a hidden Markov model to describe trajectories of hands around objects in order to help object classification in top-view videos. Similarly, [Kjellstrom et al., 2008] uses a conditional random field to detect the hand and the object for three types of actions where people grasp an object: *look through binoculars*, *drink from a cup*, *pour from a pitcher*. The work of [Stark et al., 2008] also focuses on detecting affordances for object grasping. They detect

a region where the hand interacts with an object and build a codebook of pairs of salient segments in this regions which they use at detection time to perform Hough voting for object localization. The idea of using hands also appears in [Gall et al., 2011] where they match a 3D body skeleton to a depth image and recover the 3D pose. After classifying the performed action, their method discovers and classifies the object in an unsupervised manner by analyzing the movement of hands.

One can also use more global methods relating actions performed by actors and object positions. For example, [Li and Fei-Fei, 2007] worked on images of actors in outdoor scenes and used a graphical model to both segment objects in the scene and classify the action and scene categories. Similarly, [Gupta et al., 2009] model relations between the person performing an action, the presence of manipulable objects and the scene type using a graphical model. [Yao and Fei-Fei, 2010b] go even further by relating the position of each body part to the person bounding box and the object using a Markov Random Field. Their algorithm uses an iterative method to cluster each action into sub-classes and learns the connectivity and the parameters of the MRF for each sub-class. [Desai et al., 2010] adopt another approach: their method samples a set of bounding boxes in each image and aims at assigning them to an object class, a body part or to background. They optimize an energy function where unaries encode the object's appearance and binaries encode coarse spatial compatibility between two object detections. Both terms also depend on the person pose which is detected using HOG templates specific to each action. This model has the advantage of learning good object classifiers but only coarsely describes person-object interactions and their relative positions. We explored the opposite approach in [Delaitre et al., 2011] where we used a bank of pre-trained object and body part detectors. Our body part detectors are probably more noisy than the pose templates of [Desai et al., 2010] but we designed a more accurate spatial model. Our method first mines discriminative pairs of detectors in a specific spatial configuration for each action. Then it uses the responses of such pairs

as image descriptors to perform action classification. Instead of relying on a single object detection next to a single strong person evidence as the previous work, we see person-object interactions as a collection of weak cues that must be aggregated. Another difference from the previous work is that we do not provide any information on the object location: the interactions are mined in a weakly-supervised setup.

The importance of weakly supervised methods has grown with the increasing size of datasets over time for which ground truth annotations become very costly. Whereas former methods use databases with 6 to 9 action classes and train from 180 to 560 images, the VOC2012 dataset [Everingham et al., 2012] has 11 classes with more than 4500 training images and the Stanford 40 Actions dataset [Yao et al., 2011a] has 40 action classes with 4000 training images. Beside our work, others have also developed weakly supervised methods where the location of objects is not known at training time. The method of [Fathi et al., 2011] developed for person-centric videos allows for example to automatically detect objects in daily activities by only annotating the name of objects present in each training video clip. It first uses optical flow to estimate a foreground segmentation and refines it via a graph cut based on spatio-temporal constraints before segmenting the individual objects in each image. Then, confident representative instances of objects are extracted and the labels are propagated across space and time. Finally, they train an object classifier by using those automatically localized examples. One could go one step further by classifying video clips into different activities by using the detected objects. Another method described in [Prest et al., 2011] assumes that exactly one object is involved per action. For each image, the method generates a set of candidate windows which may contain the object of interest by using the objectness measure of [Alexe et al., 2010]. Then, based on a cost function, the method selects the window which is most likely to contain the object and whose appearance and location best agrees with other images in the database. This allows them to learn an object classifier and a person-relative location prior in an

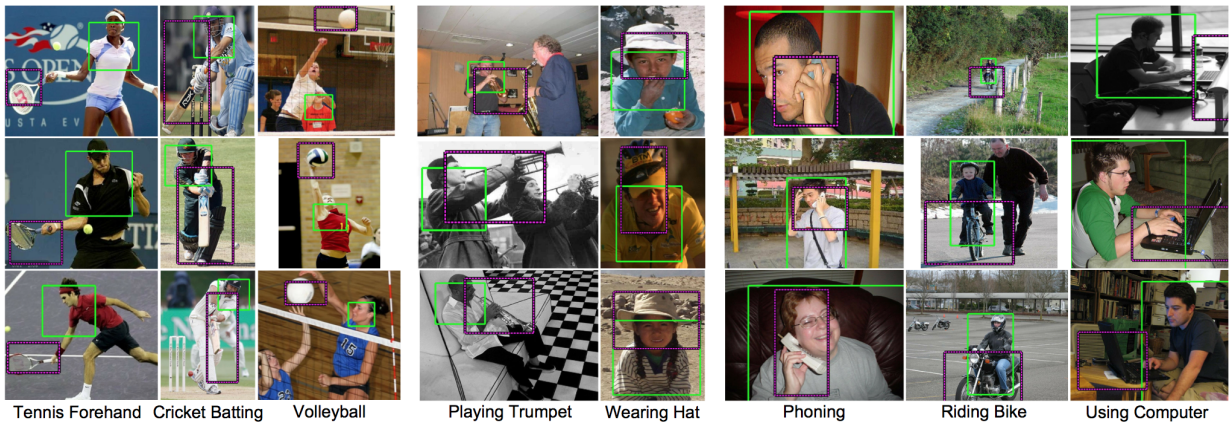


Figure 2.8 – Some automatic action-object detection examples of the model described in [Prest et al., 2011]. The green (resp. dashed purple) boxes correspond to the person (resp. the object) detections.

unsupervised manner to perform action classification in still images. See Figure 2.8 for an example of actions and detected bounding boxes.

2.5.2 People and scenes

A scene is a combination of objects and stuff (e.g. *car, buildings* or *roads* for a city, or *sofa, window* or *wall* for a living room). The spatial arrangement of objects can change significantly from scene to scene. Object detectors can be used to help scene understanding and the same is true for people. For example persons would appear and disappear through a door or sit on some specific object. We thus review next the works that models interactions between scenes and people's trajectories, pose and action.

We begin with work aiming at inferring plausible actions and interactions in scenes containing no people. For example, [Vu et al., 2014] have manually labeled different plausible actions for a set of outdoor scenes and used this dataset to perform action prediction from a static scene using a standard SVM classifier learned from Fisher Vector descriptors. Others have looked at indoor scenes, e.g.: the method of [Grabner et al., 2011] defines a non-parametric prior on relative position between an actor and a chair. It then slides a 3D

actor in a 3D scene to detect sittable surfaces. Similarly, [Gupta et al., 2011] pick sitting and lying poses using motion capture and manually define contact points with support surfaces. The method then slides the poses on the output of a geometric layout estimation algorithm to detect areas where people can sit or lay down in images, see Figure 2.9.

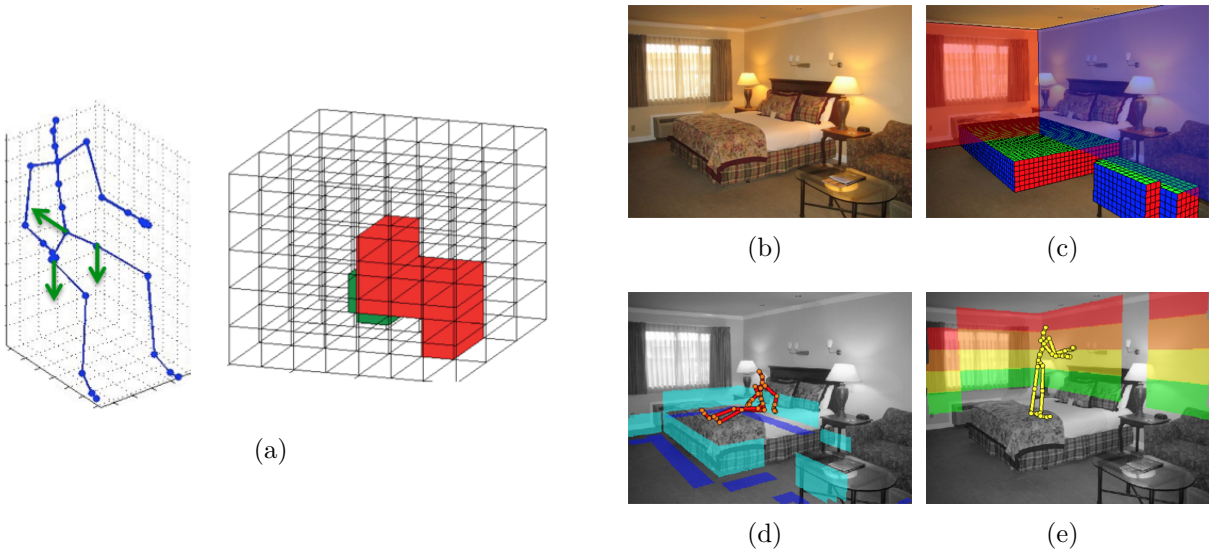


Figure 2.9 – Some plausible surfaces and poses estimated by the model of [Gupta et al., 2011]: (a) support surfaces and occupied volume of a "sitting" pose, (b) the input image, (c) estimated 3D occupancy map, (d) estimation of the possible surface for the pelvic joint (in blue) and the back location (in cyan) for a "sitting reclined" pose, (e) vertical surfaces that a person's hand can touch for a "reaching" pose, from low (green) to high (red).

The two previous methods use the free space in a 3D scene as input to output regions where people could be observed in a specific pose, e.g. sitting or lying. The following works adopt the reverse approach and use people as input sensors. They detect typical poses, infer sittable or reachable surfaces and use them to segment objects, correct the 3D layout or even approximate the free space in 3D. For example, the algorithm of [Peursum et al., 2005] recovers four objects and the floor in an office. It first segments the background into super-pixel regions and uses a Bayes classifier to label the regions. This classifier uses features generated from a pose estimation method combined with an action classifier. We have

explored similar ideas in [Fouhey et al., 2012] and [Delaitre et al., 2012] but have worked on much more challenging consumer videos downloaded from Youtube depicting real-world cluttered scenes. In [Fouhey et al., 2012], we detect standing, sitting and reaching poses and accumulate cues over time to identify the floor as well as "sittable" and "reachable" surfaces such as sofa and tables. We use the surface type labels to correct the output of the room layout estimation method of [Hedau et al., 2009] and get an approximation of the 3D free space. In [Delaitre et al., 2012], we used the same data to build an object classifier. Based on the pose of people and the image appearance, the output of this classifier serves as a cue for soft-segmentation of cluttered scenes into eleven object classes, including floor, wall and ceiling. It shows significantly better performance when person and appearance cues are combined rather than using appearance alone.

Similar ideas were also explored in outdoor videos to segment regions of interest by tracking people and objects. For example, [Wang et al., 2006a] introduce a measure of similarity between object trajectories to cluster them into groups with the similar properties. This allows them to describe a video by the stream of objects, detect the starting and ending points of trajectories and detect outliers. [Turek et al., 2010] define a descriptor of tracks of objects and use a bag-of-words aggregation to describe each region of a video by a histogram of "track words". The method uses mean-shift to cluster regions having the same functional description in an unsupervised manner, thus obtaining an automatic segmentation of the scene into roads, side walks, building, etc..

A whole group of methods is interested in doing the opposite task. They aim at forecasting moving agent trajectories from the scene layout. The work of [Yuen and Torralba, 2010] introduces a simple method to predict object trajectories in still images. They first track and cluster trajectories in training videos. Then, for a given test video, they find its nearest neighbors in the training set using GIST-based retrieval [Oliva and Torralba, 2001] and transfer the tracks from the matched scenes. The method of [Kitani et al., 2012] defines

a more sophisticated statistical model of people's motion relating their position, their velocity and the visual features extracted from grass, pavement, sidewalk, car and other object detectors. They learn the parameters of the model by maximizing its likelihood on observed training trajectories and forecast the possible trajectories starting at a given location at test time. [Walker et al., 2014] propose a similar method to predict the trajectories of automatically identified moving agents in a temporal sequence of still images. To do so they first extract a collection of mid-level patches [Doersch et al., 2012, Singh et al., 2012, Doersch et al., 2013] and track them along the image sequence. This allows them to compute the probability that a given patch type (e.g. a front facing car) representing a moving agent evolves in a certain direction or transits to another type of mid-level patch (e.g. a right facing car) between two consecutive frames. Their method also computes a compatibility prior between each agent and each super-pixel of the test image sequence by gathering nearest neighbor super-pixels from the training set and analyzing if those regions interact with the agent. Those probabilities are then used to find the shortest path between an initial and the final object-scene configuration.

Part I

Improving action classification

Chapter 3

Bag-of-features model for action classification

3.1 Introduction

In this chapter, we address the problem of classifying common human actions represented in consumer photographs. We assume that we are given a set of training still images with each person outlined by a bounding box and annotated with a specific action. The goal is to learn the parameters of a model to automatically recognize the action performed by a "test" person appearing in an image which does not belong to the training set. The outline of the person is also given by a bounding box at test time.

The existing methods [Gupta et al., 2009, Ikizler et al., 2008, Wang et al., 2006b] have mainly relied on the body pose as a cue for action recognition. While promising results have been demonstrated on sports actions [Gupta et al., 2009, Ikizler et al., 2008, Wang et al., 2006b], typical action images such as the ones illustrated in Figure 3.1 often contain heavy occlusions and significant changes in camera viewpoint and hence present a

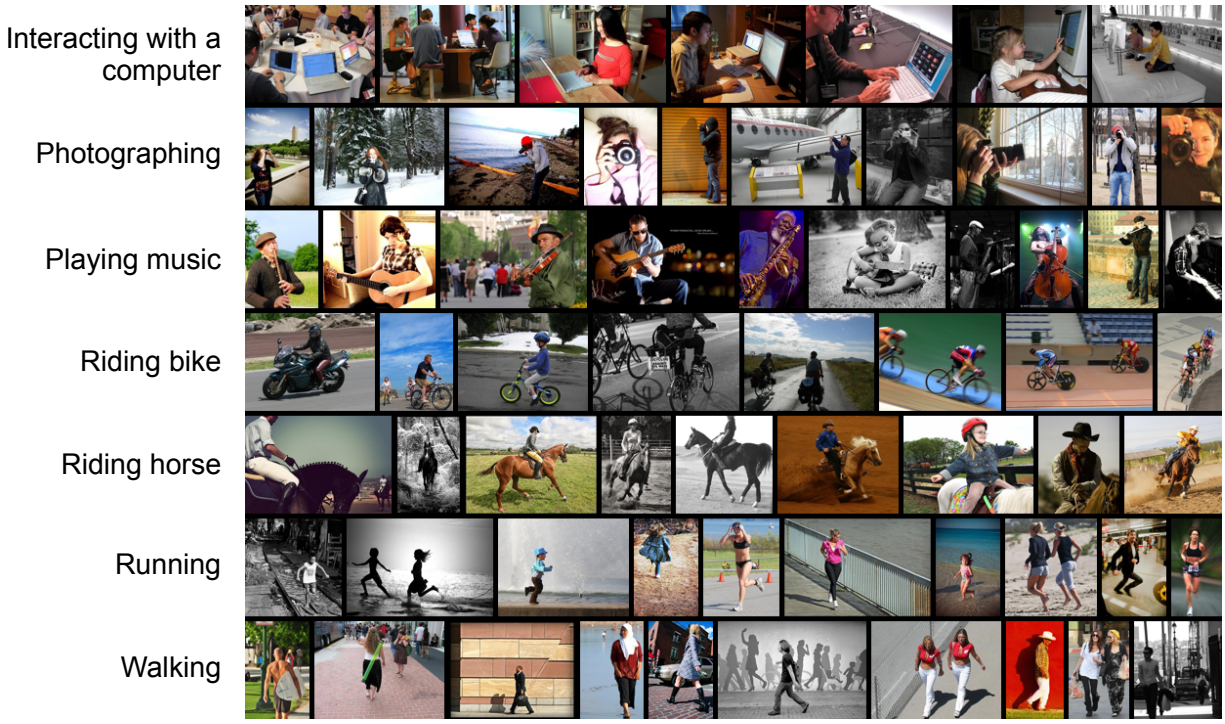


Figure 3.1 – Example images from our newly collected dataset of seven action classes. Note the natural and challenging variations in the camera view-point, clothing of people, occlusions, object appearance and the scene layout present in the consumer photographs.

serious challenge for current body-pose estimation methods. At the same time, the presence of particular objects [Gupta et al., 2009] and scene type [Li and Fei-Fei, 2007] often characterize the action and can be used for action recognition.

To deal with various types of actions in still images, we avoid explicit reasoning about body poses and investigate more general classification methods. We study the impact of scene context on action recognition in typical consumer photographs and construct a new dataset [Delaitre et al., 2010b] with seven classes of actions in 911 images obtained from the Flickr photo-sharing web-site. Image samples in Figure 3.1 illustrate the natural and challenging variations of actions in our dataset with respect to the camera view-point, clothing of people, occlusions, object appearance and the scene layout.

We study performance of statistical bag-of-features representations combined with SVM classification [Zhang et al., 2007]. In particular, we investigate person-centric representa-

tions and study the influence of the contextual information provided by the scene on action recognition. We try a large set of parameters on the validation set and show a consistent generalization of results to the test set. In addition to statistical methods, we investigate the structural part-based DPM model of [Felzenszwalb et al., 2009] reviewed in Section 2.1 and demonstrate improved performance with the combination of both models. Based on the comparative evaluation on the datasets of [Gupta et al., 2009] and [Yao and Fei-Fei, 2010a] we demonstrate that previous methods relying on explicit body-pose estimation can be significantly outperformed by more generic recognition methods investigated in this chapter.

Chapter outline: In Section 3.2 we describe our new dataset for action recognition in still images and detail performance measures used in our evaluation. Section 3.3 presents the bag-of-features (BOF) model and 3.4 focuses on the deformable part model from [Felzenszwalb et al., 2009] and its combination with the BOF model. Section 3.5 provides extensive experimental evaluation of different methods and parameter settings on the three still-image action datasets. We conclude the chapter in Section 3.6 with a discussion of the model and its possible extensions.

3.2 Datasets and performance measures

We consider three datasets in this chapter: the datasets of Gupta et al. [Gupta et al., 2009] and Yao and Fei-Fei [Yao and Fei-Fei, 2010a], focused on sports and people playing musical instruments, respectively, as well as our newly collected dataset of actions in consumer photographs available at [Delaitre et al., 2010b]. To avoid the focus on a specific domain and also investigate the effect of background (images in [Gupta et al., 2009, Yao and Fei-Fei, 2010a] are cropped to eliminate background) we collect a new dataset of full (non-cropped) consumer photographs depicting seven common human actions: “Interacting with computers”, “Photographing”, “Playing a musical instrument”, “Riding bike”,

"Riding horse", "Running" and "Walking". Images for the "Riding bike" action were taken from the Pascal 2007 VOC Challenge and the remaining images were collected from Flickr by querying on keywords such as "running people" or "playing piano". Images clearly not depicting the action of interest were manually removed. This way we have collected a total of 911 photos. Each image was manually annotated with bounding boxes indicating the locations of people. For these annotations we followed the Pascal VOC guidelines. In particular, we labeled each person with a bounding box which is the smallest rectangle containing its visible pixels. The bounding boxes are labelled as "Truncated" if more than 15%-20% of the person is occluded or lies outside the bounding box. We also added a field "action" to each bounding box to list all actions being executed. We collected at least 109 persons for each class, split into 70 persons per class for training and the remaining ones for test. Example images for each of the seven classes are shown in Figure 3.1.

Performance measures: We use two performance measures throughout the chapter: (i) the classification accuracy and (ii) the mean average precision (mAP). The classification accuracy is obtained as the average of the diagonal of the confusion table between different classes, and is a typical performance measure for multi-way classification tasks. To obtain mAP we first compute the area under the precision-recall curve (average precision) for each of the seven binary 1-vs-all action classifiers. mAP is then obtained as the mean of average precisions across the seven actions.

3.3 Bag-of-features classifier

Here we describe the spatial pyramid bag-of-features representation [Lazebnik et al., 2006] with the Support Vector Machine (SVM) classifier [Schölkopf and Smola, 2002] and the implementation choices investigated in this chapter. In particular we detail the image

representation, the different kernels of the SVM classifier, and different methods for incorporating the person bounding box and the scene background information into the classifier.

Image representation: Images (or image regions given by a rectangular bounding box) are represented using SIFT descriptors sampled on 10 regular grids with increasing scales with spacing $s_i = \lfloor 12 \cdot 1.2^i \rfloor$ pixels for $i = 0, \cdots, 9$. The scale of features extracted from each grid is set to $w_i = 0.2 \cdot s_i$. Visual vocabularies are built from training descriptors using k-means clustering. We consider vocabularies of sizes $K \in \{256, 512, 1024, 2048, 4096\}$ visual words. Descriptors from both training and test sets are then assigned to one of the visual words and aggregated into a K-dimensional histogram, denoted further as the bag-of-features representation. Following the spatial pyramid representation of Lazebnik et al. [Lazebnik et al., 2006] we further divide the image into 1×1 (Level 0), 2×2 (Level 1) and 4×4 (Level 2) spatial grids of cells. Local histograms within each cell are then concatenated with weights 0.25, 0.25 and 0.5 for levels 0, 1, and 2, respectively. This results in a $(1 + 4 + 16)K = 21K$ dimensional representation, where K is the vocabulary size. The weights of the different histogram levels are kept fixed throughout the experiments, but could be potentially learnt as shown in [Bosch et al., 2007]. This representation captures a coarse spatial layout of the image (or an image region) and has been shown beneficial for scene classification in still images [Lazebnik et al., 2006] and action classification in videos [Laptev et al., 2008].

Support vector machine classification: Classification is performed with the SVM classifier using the 1-vs-all scheme, which, in our experiments, resulted in a small but

consistent improvement over the 1-vs-1 scheme. We investigate four different kernels:

1. the histogram intersection kernel, given by $\sum_i \min(x_i, y_i)$; (3.1)
2. the χ^2 kernel, given by $\exp\left(-\frac{1}{\gamma}\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}\right)$; (3.2)
3. the Radial basis function (RBF) kernel, given by $\exp\left(-\frac{1}{\beta}\sum_i (x_i - y_i)^2\right)$; (3.3)
4. the linear kernel given by $\sum_i x_i y_i$. (3.4)

$\mathbf{x}$ and $\mathbf{y}$ denote visual word histograms, and γ and β are kernel parameters. For the χ^2 and intersection kernels, histograms are normalized to have unit L1 norm. For the RBF and linear kernels, histograms are normalized to have unit L2 norm [Vedaldi et al., 2009]. Parameters γ and β of the χ^2 and RBF kernels, respectively, together with the regularization parameter of the SVM are set for each experiment by a 5-fold cross validation on the training set.

Incorporating the person bounding box into the classifier: Previous work on object classification [Zhang et al., 2007] demonstrated that scene is often correlated with objects in the image (e.g. cars often appear on streets) and can provide useful signal for the classifier. The goal here is to investigate different ways of incorporating the scene information into the classifier for actions in still images. We consider the following four approaches, see Figure 3.2 for illustrations:

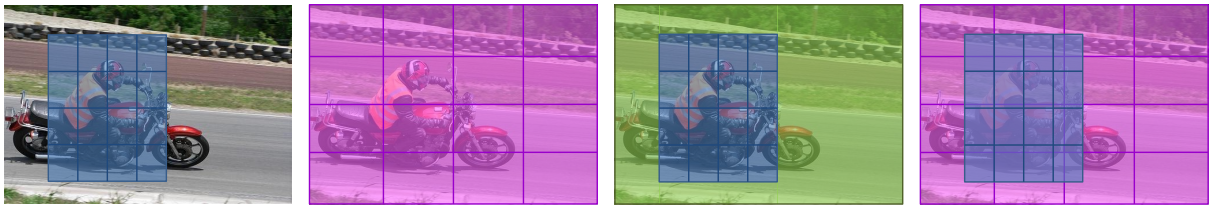
A. **"Person":** Images are centred on the person performing the action, cropped to contain 1.5× the size of the bounding box and re-sized such that the larger dimension is 300 pixels. This setup is similar to that of [Gupta et al., 2009], i.e. the person occupies the majority of the image and the background is largely suppressed.

B. **"Image":** The original images are resized to have the larger dimension at most 500 pixels. No cropping is performed. The person bounding box is not used in any stage of training or testing apart from evaluating the performance. In this setup, the visual word histograms represent a mix of the action and the scene.

C1. **"Person+Scene":** The original images are resized so that the maximum dimension of the 1.5× rescaled person bounding box is 300 pixels, but no cropping is performed. The 1.5× rescaled person bounding box is then used in both training and test to localize the person in the image and provides a coarse segmentation of the image into person (inside the rescaled person bounding box) and scene (the rest of the image). The person and scene context are treated separately. The final kernel value between two images X and Y represented using person histograms $\mathbf{x}_p$ and $\mathbf{y}_p$, and scene histograms $\mathbf{x}_s$ and $\mathbf{y}_s$, respectively, is given as the sum of the two kernels, $K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}_p, \mathbf{y}_p) + k(\mathbf{x}_s, \mathbf{y}_s)$ where k is one of the four kernels (3.1-3.4). The person region is represented using a 2-level spatial pyramid whereas the scene is represented using a BOF histogram with no spatial binning.

C2. **"Person+Image"**: This setup is similar to C1, however, instead of the scene region, 2-level spatial pyramid representation of the entire image is used.

Note that approaches A, C1 and C2 use the manually provided person bounding boxes at both the training and test time to localize the person performing the action. This simulates the case of a perfectly working person detector [Dalal and Triggs, 2005, Felzenszwalb et al., 2009].



(A) (B) (C1) (C2)

Figure 3.2 – Different experimental setups with different use of the scene context: (A) no scene context, (B) no differentiation between person and scene context, (C1) person and scene context are treated separately, no spatial pyramid is applied to the latter, (C2) person and scene context are treated separately, spatial pyramid applied to each region.

3.4 Discriminatively trained part-based model

We also investigate the performance of the discriminatively trained part-based model of [Felzenszwalb et al., 2009] (DPM), which, in contrast to the bag-of-features approach, provides a deformable part-based representation of each action. The approach combines the strengths of efficient pictorial structure models [Felzenszwalb and Huttenlocher, 2005, Fischler and Elschlager, 1973] with recent advances in discriminative learning of SVMs with latent variables [Yu and Joachims, 2009]. The approach has shown excellent human and object detection performance in the PASCAL VOC challenge [Everingham et al., 2007].

In this chapter we apply the model for classification (rather than detection with spatial localization) and focus on recognition of human actions rather than objects. Actions are modeled as multi-scale HOG templates with flexible parts. Similarly to the spatial pyramid bag-of-features representation described in Section 3.3, we train one model for each action class in a 1-vs-all fashion. Positive training data is given by the 1.5× rescaled person bounding boxes for the particular action and negative training data is formed from all images of the other action classes. At test time, we take the detection with the maximum score, which overlaps the manually specified person bounding box in the test image more than 50%. The overlap is measured using the standard ratio of areas of the intersection

over the union. The 50% overlap allows for some amount of scale variation between the model and the manual person bounding box. In cases when the person bounding box is not available the detection with the maximum score over the entire image is taken. We use the version 4 of the training and detection code available at [Felzenszwalb et al., 2008a], which supports models with multiple mixture components for each part allowing for a wider range of appearances of each action. We train models with 8 parts and 3 mixture components.

Combining the part-based model with the bag-of-features classifier: The part-based model (DPM) represents mostly the person and its immediate surroundings and largely ignores the background information. Hence, we also investigate combining the model with bag-of-feature classifiers described in Section 3.3. We demonstrate in Section 3.5 that such combination can significantly improve the classification performance of the DPM approach. The two approaches are combined by simply adding together their classification scores with equal weighting. However, the weights could be potentially learned. In a similar fashion, combining scene-level classifiers with object detectors was shown to improve object detection results in the PASCAL 2009 object detection challenge [Harzallah et al., 2009].

3.5 Results

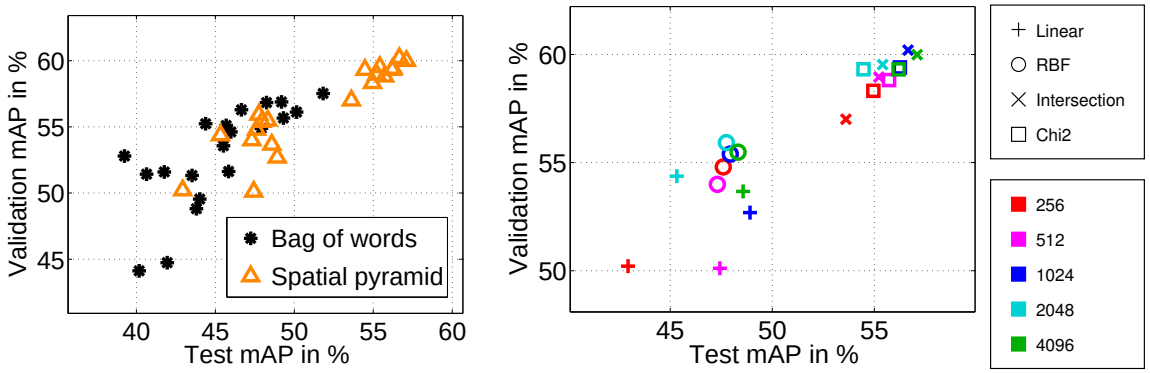
We first evaluate different parameter settings for the bag-of-features classifier. Equipped with a well tuned classifier we examine different ways of incorporating the foreground (person) and background (scene context) information. Next, we compare and combine the bag-of-features classifier with the structured part-based model. Finally, we show results on the datasets of [Gupta et al., 2009] and [Yao and Fei-Fei, 2010a].

Setting parameters for the bag-of-features method: We first evaluate in detail different parameter settings (kernel type, vocabulary size, spatial representation) for bag-of-features method A, where images are cropped to contain mostly the person performing the action and the background is suppressed. We have found that the pattern of results across different parameter settings for methods B and C is similar to A and hence their detailed discussion is omitted.

Figure 3.3 shows plots of the classification performance obtained from the 5-fold cross-validation on the training set against the classification performance on the test set. First, we note that both cross-validation and test performance are well correlated, which suggests that the cross-validation results can be used to select the appropriate parameter setting. It is clear from Figure 3.3(a) that spatial pyramid representation outperforms the vanilla bag-of-features model with no spatial binning. Examining Figure 3.3(b), all kernels show similar trend of improvement towards larger vocabulary sizes. However the χ^2 and intersection kernels convincingly outperform the linear and RBF kernels. The best results (in terms of the lowest cross-validation error) are obtained for the spatial pyramid representation, intersection kernel, and vocabulary size 1,024 and we use this parameter setting for the rest of the chapter.

How to model scene context? Here we examine the different approaches for incorporating the scene information into the bag-of-features action classifier (methods A-C). The overall results are summarized using the classification accuracy and the mean average precision in table 3.1 (rows A-C2). Average precision across different action classes is shown in table 3.2 (columns A-C2).

Focusing on the person by cropping the image and removing the background (method A) results in a slightly better overall performance than method B where we use the entire image, including the background, with no knowledge about the location of the person.



(a) (b)

Figure 3.3 – Classification performance (cross-validation mAP vs. test mAP) for different parameter settings for the BOF method "A. Person". The best results are at the top right portion of the graph. (a) Spatial pyramid vs. the bag-of-feature representation. (b) Classification performance for different combinations of kernels and vocabulary sizes using the spatial pyramid representation. Best viewed in color. The standard deviation (not shown in the plots) of the validation mAP is typically 2-3%.

However, for some actions ("Playing Music" and "Riding Bike") using the background (method B) is beneficial and reduces their confusion with other classes.

The overall performance can be further improved by treating and matching the person and scene regions separately using two separate kernels (methods C1 and C2). This holds for all classes except "Running" and "Walking" where using the scene (method C2) slightly increases the confusion with the other action classes compared to method A (and specially with "Riding Bike" and "Riding Horse" which are also outdoor scenes). In addition, representing the scene with a spatial pyramid (C2) performs better overall than the vanilla BOF histogram (C1) with no spatial information. The overall benefit of treating person and scene regions separately is inline with the experimental evidence from object and image classification [Uijlings et al., 2009, Zhang et al., 2007].

Part-based model vs. bag-of-features classifier: Here we compare the performance of the bag-of-features classification method (C2), the structured part-based model (DPM) and their combination (DPM+C2). The overall results are summarized using the classi-

	Method	mAP	Accuracy
A.	BOF Person	56.7	55.9
B.	BOF Image	54.0	54.0
C1.	BOF Person+Background	57.6	56.8
C2.	BOF Person+Image	59.6	58.9
	DPM	50.2	55.1
	DPM + C2	**62.9**	**62.2**

Table 3.1 – The overall classification performance for the different methods.

Action / Method	A	B	C1	C2	DPM	DPM+C2
(1) Inter. w/ Comp.	51.6	51.6	57.3	58.2	30.2	**58.5**
(2) Photographing	31.8	30.7	24.7	35.4	28.1	**37.4**
(3) Playing Music	63.4	69.1	68.4	**73.2**	56.3	73.1
(4) Riding Bike	76.5	78.2	82.3	82.4	68.7	**83.3**
(5) Riding Horse	66.2	56.3	67.4	69.6	60.1	**77.0**
(6) Running	51.3	39.0	45.1	44.5	52.0	**53.3**
(7) Walking	55.8	53.2	**58.0**	54.2	56.0	57.5
mAP	56.7	54.0	57.6	59.6	50.2	**62.9**

Table 3.2 – Per-class average precision across different methods.

fication accuracy and mean average precision in the last three rows of table 3.1. Average precision across different action classes is shown in the last three columns of table 3.2. The bag-of-features classifier (C2) and structured part-based model (DPM) have comparable accuracy but the average precision of (C2) is better for all classes but "Running" and "Walking". It might be due to our choice to limit the part-based model to 3 mixture components: this could be insufficient for classes as "Interacting with computer", "Photographing" or "Playing music" where there is a large number of viewpoints (including close-ups) and interacting objects. Overall, the combined (DPM+C2) approach performs best and significantly improves over (C2) on classes like "Running" and "Walking" where (DPM) was better, but also interestingly on "Riding horse" which suggests that the bag-of-features classifier and the part-based model use complementary information for this class. These variations across classes are likely due to the varying levels of consistency of the hu-

man pose (captured well by structured part-based models), and the overall scene (captured well by the bag-of-features classifier). The full confusion table for the overall best performing method (DPM+C2) is shown in table 3.3. While accuracy is over 65% on actions like "Interacting with computer", "Playing music", "Riding bike" or "Riding horse" other actions are more challenging, e.g. "Photographing" (accuracy 30%) is often confused with actions where people mostly stand as "Playing music", "Running", or "Walking". The last two classes have accuracy around 55% and are often confused with each other. Examples of images correctly classified by the combined DPM+C2 method are shown in figures 3.4 and 3.5. Examples of challenging images misclassified by the DPM+C2 method are shown in Figure 3.6. We have found that the combined DPM+C2 method often improves the output of the bag-of-features classifier (C2) on images with confusing (blurred, textureless or unusual) background, but where the pose of the person is very clear and the DPM model provides a confident output. Similarly, the combined method appears to improve the vanilla DPM results mainly in cases where camera viewpoint or the pose of the person are unusual.

Action	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Inter. w/ Comp.	**82.1**	0.0	15.4	2.7	0.0	0.0	0.0
(2) Photographing	13.0	**29.9**	15.6	3.9	6.5	13.0	18.2
(3) Playing Music	12.7	15.3	**66.1**	2.5	2.5	0.0	0.9
(4) Riding Bike	0.0	2.9	5.0	**74.1**	7.9	2.2	7.9
(5) Riding Horse	0.0	0.0	7.0	12.3	**73.7**	1.8	5.3
(6) Running	2.5	6.2	7.4	8.6	0.0	**51.9**	23.5
(7) Walking	4.9	8.2	0.0	0.0	11.5	18.0	**57.4**

Table 3.3 – Confusion table for the best performing method (DPM+C2). Accuracy (average of the diagonal): 62.2%.

Comparison with the methods of [Gupta et al., 2009], [Yao and Fei-Fei, 2010a] and [Yao and Fei-Fei, 2010b]: For the sport dataset of [Gupta et al., 2009] we have cross-validated again the parameters of the bag-of-features classifier and found that bigger

DPM+C2: | RidingBike	PlayingMusic	Photographing	Running	Walking
C2: | RidingHorse	Inter. w/ comp.	RidingBike	Inter. w/ comp.	Running

Figure 3.4 – Example images correctly classified by the combined DPM+C2 method (labels in the 2nd row), but misclassified by the C2 bag-of-features approach (labels in the 3rd row).

DPM+C2: | Photographing	Inter. w/ comp.	RidingHorse	RidingBike	PlayingMusic
DPM: | PlayingMusic	PlayingMusic	RidingBike	RidingHorse	Photographing

Figure 3.5 – Example images correctly classified by the combined DPM+C2 method (labels in the 2nd row), but misclassified by the part-based DPM approach (labels in the 3rd row).

DPM+C2: | Walking	Running	Photographing	Photographing	PlayingMusic
G.T.: | Running	Photographing	PlayingMusic	Walking	Inter. w/ comp.

Figure 3.6 – Examples of challenging images misclassified by the combined DPM+C2 method (labels in the 2nd row). The ground truth labels are shown in the 3rd row. Note the variation in viewpoint and scale as well as partial occlusion.

Dataset	[Gupta et al., 2009]		[Yao and Fei-Fei, 2010a]			
			Task 1		Task 2	
Method	mAP	Acc.	mAP	Acc.	mAP	Acc.
[Gupta et al., 2009]	–	78.7	–	–	–	–
[Yao and Fei-Fei, 2010a]	–	–	–	65.7	–	80.9
[Yao and Fei-Fei, 2010b]	–	83.3	–	–	–	–
BOF Image (B)	91.3	**85.0**	76.9	71.7	87.7	83.7
DPM	77.2	73.3	53.6	67.6	82.2	82.9
DPM + BOF Image (B)	**91.6**	**85.0**	**77.8**	**75.1**	**90.5**	**84.9**

Table 3.4 – Comparison with the method of [Gupta et al., 2009] and of [Yao and Fei-Fei, 2010a, Yao and Fei-Fei, 2010b] on their datasets. 'Task 1' is the 7-class classification problem and 'Task 2' is the PPMI+ vs PPMI- problem (see [Yao and Fei-Fei, 2010a]).

vocabularies ($K = 4096$) perform better on this dataset. Other parameters (the intersection kernel and spatial pyramid binning) remain the same. For the Person Playing Musical Instrument dataset of [Yao and Fei-Fei, 2010a] we adopted a denser sampling of the SIFT features with initial spacing of 6 pixels to adapt to the smaller size of the images. Moreover we used a 3 level spatial pyramid and a DPM with 9 parts to have a denser spatial coverage. Other parameters (the intersection kernel and $K = 1024$) remain the same. For both datasets, no person bounding box information is used in training or test (method B). However, as the images in the original dataset are already cropped and centred to contain mostly the person of interest the approach is comparable with method A on our dataset. As shown in table 3.4, both the BOF and DPM+BOF methods outperform the approach of Gupta et al. and Yao and Fei-Fei by 1.7% to 9.4%.

3.6 Discussion

We have studied the performance of the bag-of-features classifier and the latent SVM model [Felzenszwalb et al., 2009] on the task of action recognition in still images. We have collected a new challenging dataset of more than 900 consumer photographs depicting

seven everyday human actions. We have demonstrated on this data, as well as two existing datasets of person-object interactions [Gupta et al., 2009, Yao and Fei-Fei, 2010a], that (i) combining statistical and structured part-based representations and (ii) incorporating scene context can lead to significant improvements in action recognition performance in still images. Almost all tested methods (except the image-level classifier B) use the manually provided person bounding boxes. One possible extension would be to incorporate real person detections, as reviewed in Section 2.1, to perform action detection. In the next chapter, we will focus on interaction between people and manipulable objects to further improve action classification performance.

Chapter 4

Learning person-object interactions

4.1 Introduction

The previous chapter demonstrated that the use of contextual information extracted from the scene can help classification of actions. As a given action tends to be exhibited by a set of typical poses and interactions between a person and an object, our goal is now to investigate discriminatively trained models of interactions between objects and human body parts. We build on the bag-of-features [Sivic and Zisserman, 2003, Csurka et al., 2004] and spatial pyramids models [Lazebnik et al., 2006], which have demonstrated excellent performance on a range of scene [Lazebnik et al., 2006], object [Harzallah et al., 2009, Vedaldi et al., 2009, Zhang et al., 2007] and action [Delaitre et al., 2010a] recognition tasks. Rather than relying on accurate estimation of body part configurations or accurate object detection in the image, we represent human actions as locally orderless distributions over body parts and objects together with their interactions. By opportunistically learning class-specific object and body part interactions (e.g. relative configuration of leg and horse detections for the riding horse action, see Figure 4.1), we avoid the extremely challenging

task of estimating the full body configuration. Towards this goal, we consider the following challenges: (i) what should be the representation of object and body part appearance; (ii) how to model object and human body part interactions; and (iii) how to choose suitable interaction pairs in the huge space of all possible combinations and relative configurations of objects and body parts.

To address these challenges, we introduce the following three contributions. First, we replace the quantized HOG/SIFT features, typically used in bag-of-features models [Delaitre et al., 2010a, Lazebnik et al., 2006, Vedaldi et al., 2009] with powerful, discriminatively trained, object and human body part detectors [Bourdev and Malik, 2009, Johnson and Everingham, 2011]. This significantly enhances generalization over appearance variation, due to e.g. clothing or viewpoint while providing a reliable signal on part locations. Second, we develop a part interaction representation, capturing pair-wise relative position and scale between object/body parts, and include this representation in a scale-space spatial pyramid model. Third, rather than choosing interacting parts manually, we select them in a discriminative fashion. Suitable pair-wise interactions are first chosen from a large pool of hundreds of thousands of candidate interactions using a linear support vector machine (SVM) with a sparsity inducing regularizer. The selected interaction features are then input into a final, more computationally expensive, non-linear SVM classifier based on the locally orderless spatial pyramid representation.

Chapter outline: We first describe our interaction model and image representation in Section 4.2. We detail the learning process and action classifier in Section4.3. In Section 4.4, we compare the proposed interaction model with the strong bag-of-features and deformable part models on the dataset of [Delaitre et al., 2010a]. We then evaluate the final model on the PASCAL VOC 10 Challenge [Everingham et al., 2010] and show competitive results with the state of the art. Finally, we summarize our contribution and its possible extensions in Section 4.5.

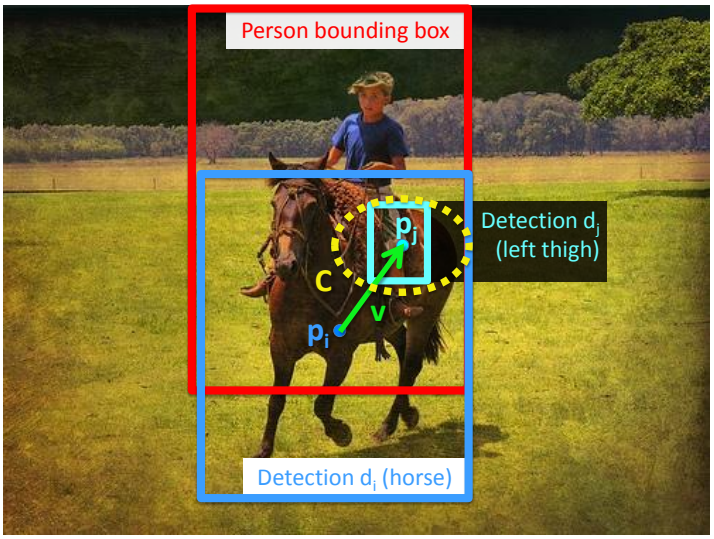


Figure 4.1 – **Representing person-object interactions** by pairs of body part (cyan) and object (blue) detectors. To get a strong interaction response, the pair of detectors (here visualized at positions $\mathbf{p_i}$ and $\mathbf{p_j}$) must fire in a particular relative 3D scale-space displacement (given by the vector $\mathbf{v}$) with a scale-space displacement uncertainty (deformation cost) given by diagonal 3×3 covariance matrix $\mathbf{C}$ (the spatial part of $\mathbf{C}$ is visualized as a yellow dotted ellipse). Our image representation is defined by the max-pooling of interaction responses over the whole image, solved efficiently by the distance transform.

4.2 Representing person-object interactions

This section describes our image representation in terms of body parts, objects and interactions among them.

4.2.1 Representing body parts and objects

We assume to have a set of n available detectors $d_1, \ldots, d_n$ which have been pre-trained for different body parts and object classes. Each detector i produces a map of dense 3D responses $d_i(\mathbf{I}, \mathbf{p})$ over locations and scales of a given image $\mathbf{I}$. We express the positions of detections $\mathbf{p}$ in terms of scale-space coordinates $\mathbf{p} = (x, y, \sigma)$ where (x, y) corresponds to the spatial location and $\sigma = \log \tilde{\sigma}$ is an additive scale parameter log-related to the image scale factor $\tilde{\sigma}$ making the addition in the position vector space meaningful.

In this chapter we use two types of detectors. For objects we use DPM detector trained on PASCAL VOC images for ten object classes[1] [Felzenszwalb et al., 2009]. For body parts we implement the method of [Johnson and Everingham, 2011] and train ten body part detectors[2] for each of sixteen pose clusters giving 160 body part detectors in total (see [Johnson and Everingham, 2011] for further details). Both of our detectors use Histograms of Oriented Gradients (HOG) [Dalal and Triggs, 2005] as an underlying low-level image representation.

4.2.2 Representing pairwise interactions

We define interactions by the pairs of detectors (d_i, d_j) as well as by the spatial and scale relations among them. Each pair of detectors constitutes a two-node tree where the position and the scale of the leaf are related to the root by scale-space offset and a spatial deformation cost. More precisely, an interaction pair is defined by a quadruplet $\mathbf{q} = (i, j, \mathbf{v}, \mathbf{C}) \in \mathbb{N} \times \mathbb{N} \times \mathbb{R}^3 \times \mathbb{M}_{3,3}$ where i and j are the indices of the detectors at the root and leaf, $\mathbf{v}$ is the offset of the leaf relatively to the root and $\mathbf{C}$ is a 3×3 diagonal matrix defining the displacement cost of the leaf with respect to its expected position. Figure 4.1 illustrates an example of an interaction between a horse and the left thigh for the horse riding action.

We measure the response of the interaction $\mathbf{q}$ located at the root position $\mathbf{p_1}$ by:

$$r(\mathbf{I}, \mathbf{q}, \mathbf{p_1}) = \max_{\mathbf{p_2}} \Big(d_i(\mathbf{I}, \mathbf{p_1}) + d_j(\mathbf{I}, \mathbf{p_2}) - \mathbf{u}^T \mathbf{C} \mathbf{u} \Big) \tag{4.1}$$

where $\mathbf{u} = \mathbf{p_2} - (\mathbf{p_1} + \mathbf{v})$ is the displacement vector corresponding to the relative position of the leaf node with respect to its expected position $(\mathbf{p_1} + \mathbf{v})$. Maximizing over $\mathbf{p_2}$ in (4.1)

[1] The ten object detectors correspond to object classes bicycle, car, chair, cow, dining table, horse, motorbike, person, sofa, tv/monitor

[2] The ten body part detectors correspond to head, torso, {left, right} × {forearm, upper arm, lower leg, thigh}

provides localization of the leaf node with the optimal trade-off between the detector score and the displacement cost. For any interaction $\mathbf{q}$ we compute its responses for all pairs of node positions $\mathbf{p_1}, \mathbf{p_2}$. We do this efficiently in linear time with respect to $\mathbf{p}$ using distance transform [Felzenszwalb and Huttenlocher, 2004].

4.2.3 Representing images by response vectors of pair-wise interactions

Given a set of M interaction pairs $\mathbf{q}_1, \cdots, \mathbf{q}_M$, we wish to aggregate their responses (4.1), over an image region $\mathcal{A}$. Here $\mathcal{A}$ can be (i) an (extended) person bounding box, as used for selecting discriminative interaction features (Section 4.3.2) or (ii) a cell of the scale-space pyramid representation, as used in the final non-linear classifier (Section 4.3.3). We define score $s(\mathbf{I}, \mathbf{q}, \mathcal{A})$ of an interaction pair $\mathbf{q}$ within $\mathcal{A}$ of an image $\mathbf{I}$ by max-pooling, i.e. as the maximum response of the interaction pair within $\mathcal{A}$:

$$s(\mathbf{I}, \mathbf{q}, \mathcal{A}) = \max_{\mathbf{p} \in \mathcal{A}} r(\mathbf{I}, \mathbf{q}, \mathbf{p}). \tag{4.2}$$

An image region $\mathcal{A}$ is then represented by a M-vector of interaction pair scores

$$\mathbf{z} = (s_1, \cdots, s_M) \text{ with } s_i = s(\mathbf{I}, \mathbf{q}_i, \mathcal{A}). \tag{4.3}$$

4.3 Learning person-object interactions

Given object and body part interaction pairs $\mathbf{q}$ introduced in the previous section, we wish to use them for action classification in still images. A brute-force approach of analyzing all possible interactions, however, is computationally prohibitive since the space of all possible interactions is combinatorial in the number of detectors and scale-space relations among

them. To address this problem, we aim in this chapter to select a set of M action-specific interaction pairs $\mathbf{q}_1, \ldots, \mathbf{q}_M$, which are both representative and discriminative for a given action class. Our learning procedure consists of the three main steps as follows. First, for each action we generate a large pool of candidate interactions, each comprising a pair of body part and/or object detectors and their relative scale-space displacement. This step is data-driven and selects candidate detection pairs which frequently occur for a particular action in a consistent relative scale-space configuration. Next, from this initial pool of candidate interactions we select a set of M discriminative interactions which best separate the particular action class from other classes in our training set. This is achieved using a linear Support Vector Machine (SVM) classifier with a sparsity inducing regularizer. Finally, the discriminative interactions are combined across classes and used as interaction features in our final non-linear spatial-pyramid like SVM classifier. The three steps are detailed below.

4.3.1 Generating a candidate pool of interaction pairs

To initialize our model, we first generate a large pool of candidate interactions in a data-driven manner. Following the suggestion in [Felzenszwalb et al., 2009] that the accurate selection of the deformation cost $\mathbf{C}$ may not be that important, we set $\mathbf{C}$ to a reasonable fixed value for all pairs, and focus on finding clusters of frequently co-occurring detectors (d_i, d_j) in specific relative configurations.

For each detector i and an image $\mathbf{I}$, we first collect a set of positions of all positive detector responses $\mathbf{P}_i^{\mathbf{I}} = \{\mathbf{p} \mid d_i(\mathbf{I}, \mathbf{p}) > 0\}$, where $d_i(\mathbf{I}, \mathbf{p})$ is the response of detector i at position $\mathbf{p}$ in image $\mathbf{I}$. We then apply a standard non-maxima suppression (NMS) step to eliminate multiple responses of a detector in local image neighbourhoods and then limit $\mathbf{P}_i^{\mathbf{I}}$ to the L top-scoring detections. The intuition behind this step is that a part/object interaction is not likely to occur many times in an image.

For each pair of detectors (d_i, d_j) we then gather relative displacements between their detections from all the training images $\mathbf{I}_k$: $\mathbf{D}_{ij} = \bigcup_k \{p_j - p_i \mid p_i \in \mathbf{P}_i^{\mathbf{I}_k} \text{ and } p_j \in \mathbf{P}_j^{\mathbf{I}_k}\}$. To discover potentially interesting interaction pairs, we perform a mean-shift clustering over $\mathbf{D}_{ij}$ using a window of radius $\mathbf{R} \in \mathbb{R}_3$ (2D-image space and scale) equal to the inverse of the square root of the deformation cost: $\mathbf{R} = diag(\mathbf{C}^{-\frac{1}{2}})$. We also discard clusters which contribute to less than η percent of the training images. The set of m resulting candidate pairs $(i, j, \mathbf{v}_1, \mathbf{C}), \cdots, (i, j, \mathbf{v}_m, \mathbf{C})$ is built from the centers $\mathbf{v}_1, \cdots, \mathbf{v}_m$ of the remaining clusters. By applying this procedure to all pairs of detectors, we generate a large pool (hundreds of thousands) of potentially interesting candidate interactions.

4.3.2 Discriminative selection of interaction pairs

The initialization described above produces a large number of candidate interactions. Many of them, however, may not be informative resulting in unnecessary computational load at the training and classification times. For this reason we wish to select a smaller number of M discriminative interactions.

Given a set of N training images, each represented by an interaction response vector $\mathbf{z_i}$, described in Equation (4.3) where $\mathcal{A}$ is the extended person bounding box given for each image, and a binary label y_i (in a 1-vs-all setup for each class), the learning problem for each action class can be formulated using the binary SVM cost function:

$$J(\mathbf{w}, b) = \lambda \sum_{i=1}^{N} \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{z_i} + b)\}^2 + \|\mathbf{w}\|_1, \tag{4.4}$$

where $\mathbf{w}, b$ are parameters of the classifier and λ is the weighting factor between the hinge loss on the training examples and the L_1 regularizer of the classifier.

By minimizing (4.4) in a one-versus-all setting for each action class we search (by binary search) for the value of the regularization parameter λ resulting in the sparse weight

vector $\mathbf{w}$ with M non-zero elements. Selection of M interaction pairs corresponding to non-zero elements of $\mathbf{w}$ gives M most discriminative (according to (4.4)) interaction pairs per action class. Note that other discriminative feature selection strategies such as boosting [Freund and Schapire, 1997] can be also used. However, the proposed approach is able to jointly search the entire set of candidate feature pairs by minimizing a convex cost given in (4.4), whereas boosting implements a greedy feature selection procedure, which may be sub-optimal.

4.3.3 Using interaction pairs for classification

Given a set of M discriminative interactions for each action class obtained as described above, we wish to train a final non-linear action classifier. We use spatial pyramid-like representation [Lazebnik et al., 2006], aggregating responses in each cell of the pyramid using max-pooling as described by equation (4.2), where $\mathcal{A}$ is one cell of the spatial pyramid. We extend the standard 2D pyramid representation to scale-space resulting in a 3D pyramid with $D = 1 + 2^3 + 4^3 = 73$ cells. Using the scale-space pyramid with D cells, we represent each image by concatenating M features from each of the K classes into a MKD-dimensional vector. We train a non-linear SVM with RBF kernel and L_2 regularizer for each action class using a 5-fold cross-validation for the regularization and kernel band-width parameters. We found that using this final non-linear classifier consistently improves classification performance over the linear SVM given by equation (4.4). Note that feature selection (Section 4.3.2) is necessary in this case as applying the non-linear spatial pyramid classifier on the entire pool of all candidate interactions would be computationally infeasible.

4.4 Experiments

We test our model on the Willow-action dataset from [Delaitre et al., 2010b] and the PASCAL VOC 2010 action classification dataset [Everingham et al., 2010]. As described in the previous chapter, the Willow-action dataset contains more than 900 images with more than 1100 labelled person detections from 7 human action classes: *Interaction with Computer, Photographing, Playing Music, Riding Bike, Riding Horse, Running* and *Walking*. The training set contains 70 examples of each action class and the rest (at least 39 examples per class) is left for testing. The PASCAL VOC 2010 dataset contains the 7 above classes together with 2 other actions: *Phoning* and *Reading*. It contains a similar number of images. Each training and testing image in both datasets is annotated with the smallest bounding box containing each person and by the performed action(s). We follow the same experimental setup for both datasets.

Implementation details: We use our implementation of body part detectors described in [Johnson and Everingham, 2011] with 16 pose clusters trained on the publicly available 2000 images database [Johnson, 2010], and 10 pre-trained PASCAL 2007 Latent SVM object detectors [Felzenszwalb et al., 2008a]: *bicycle, car, chair, cow, dining table, horse, motorbike, person, sofa, tvmonitor*. In the human action training/test data, we extend each given person bounding box by 50% and resize the image so that the bounding box has a maximum size of 300 pixels. We run the detectors over the transformed bounding boxes and consider the image scales $s_k = 2^{k/10}$ for $k \in \{-10, \cdots, 10\}$. At each scale we extract the detector response every 4 pixels and 8 pixels for the body part and object detectors, respectively. The outputs of each detector are then normalized by subtracting the mean of maximum responses within the training bounding boxes and then normalizing the variance to 1. We generate the candidate interaction pairs by taking the mean-shift

radius $\mathbf{R} = (30, 30, \log(2)/2)$, $L = 3$ and $\eta = 8\%$. The covariance of the pair deformation cost $\mathbf{C}$ is fixed in all experiments to $\mathbf{R}^{-2}$. We select $M = 310$ discriminative interaction pairs to compute the final spatial pyramid representation of each image.

Results: Table 4.1 summarizes per-class action classification results (reported using average precision for each class) for the proposed method (d. Interactions), and three baselines. The first baseline (a. BOF) is the bag-of-features classifier [Delaitre et al., 2010a], aggregating quantized responses of densely sampled HOG features in spatial pyramid representation, using a (non-linear) intersection kernel. Note that this is a strong baseline, which was shown [Delaitre et al., 2010a] to outperform the person-object interaction models of [Yao and Fei-Fei, 2010a] and [Gupta et al., 2009] on their own datasets. The second baseline (b. DPM) is the latent SVM classifier [Felzenszwalb et al., 2009] trained in a 1-vs-all fashion for each class. To obtain a single classification score for each person bounding box, we take the maximum DPM detection score from the detections overlapping the extended bounding box with the standard overlap score [Everingham et al., 2010] higher than 0.5. The final baseline (c. Detectors) is a SVM classifier with an RBF kernel trained on max-pooled responses of the entire bank of body part and object detectors in a spatial pyramid representation but without interactions. This baseline is similar in spirit to the object bank representation [Li et al., 2010], but here targeted to action classification by including a bank of pose-specific body part detectors as well as object detectors. On average, the proposed method (d.) outperforms all baselines, obtaining the best result on 4 out of 7 classes. The largest improvements are obtained on Riding Bike and Horse actions, for which reliable object detectors are available. The improvement of the proposed method d. with respect to using the plain bank of object and body part detectors c. directly demonstrates the benefit of modeling interactions. Example detections of interaction pairs are shown in Figure 4.2.

Action / Method	a. BOF	b. DPM	c. Detectors	d. Interactions
(1) Inter. w/ Comp.	**58.2**	30.2	45.6	56.6
(2) Photographing	35.4	28.1	36.4	**37.5**
(3) Playing Music	**73.2**	56.3	68.4	72.0
(4) Riding Bike	82.4	68.7	86.7	**90.5**
(5) Riding Horse	69.6	60.1	71.4	**75.0**
(6) Running	44.5	52.0	57.7	**59.7**
(7) Walking	54.2	56.0	**57.7**	57.6
Average (mAP)	59.6	50.2	60.5	**64.1**

Table 4.1 – Per-class average-precision for different methods on the Willow-actions dataset. See text for description.

Table 4.2 shows the performance of the proposed interaction model (d. Interactions) and its combination with the baselines (e. BOF+DPM+Inter.) on the Pascal VOC 2010 data. Interestingly, the proposed approach is complementary to both the BOF (51.3 mAP) and DPM (44.1 mAP) methods and by combining all three approaches (following [Delaitre et al., 2010a]) the overall performance improves to 60.7 mAP. We also report results of the "Poselet" method [Maji et al., 2011], which, similar to our method, is trained from external non-Pascal data. Our combined approach achieves better overall performance and also outperforms the "Poselet" approach on 6 out of 9 classes. Finally, our combined approach also obtains competitive performance compared to the overall best reported competition result on the Pascal VOC 2010 data – "SURREY_MK_KDA", details available on [Everingham et al., 2010] – and outperforms this method on the "Riding Horse", "Taking Photo" and "Walking" classes.

4.5 Discussion

We have developed person-object interaction features based on non-rigid relative scale-space displacement of pairs of body part and object detectors. Further, we have shown that such features can be learned in a discriminative fashion and can improve action clas-

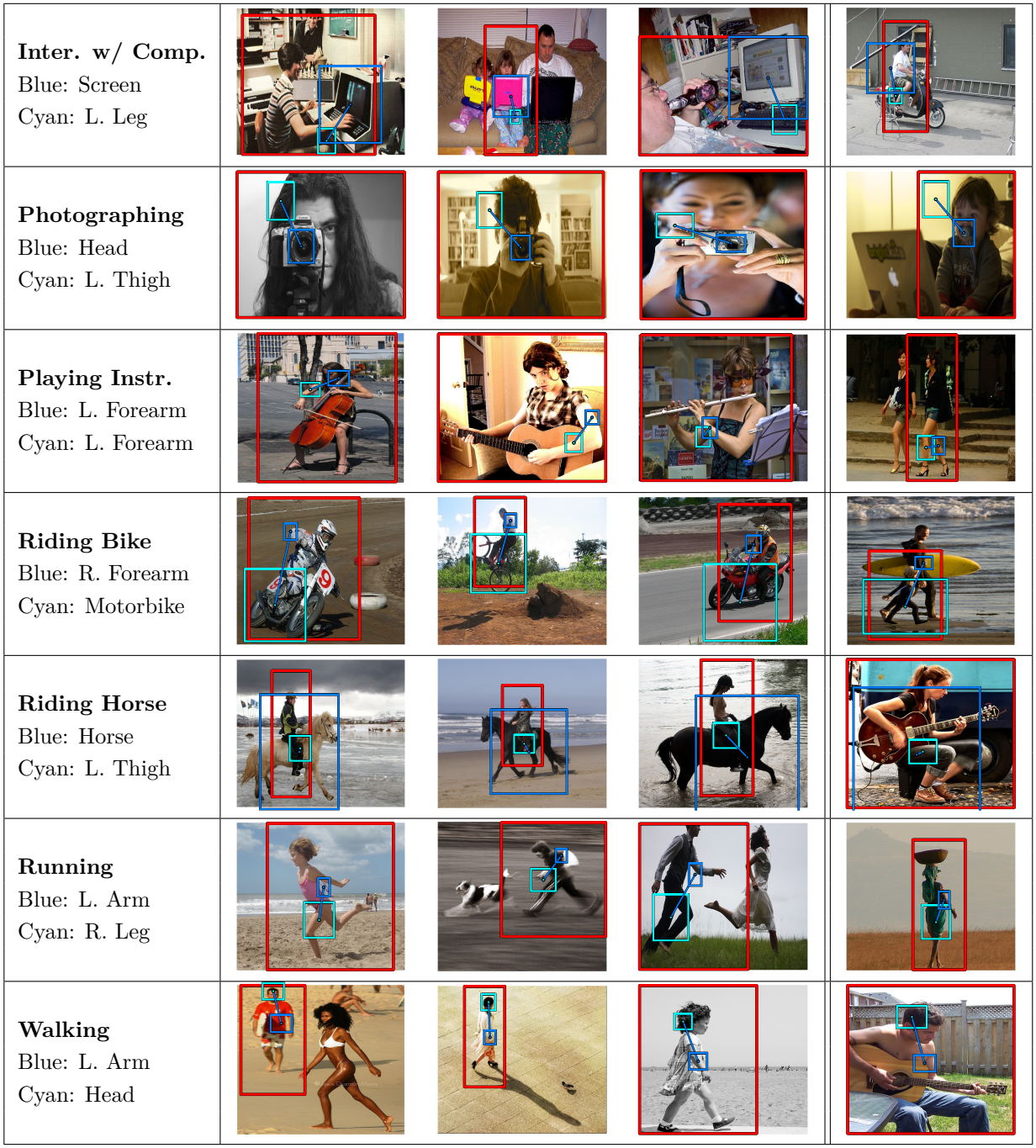


Figure 4.2 – **Example detections of discriminative interaction pairs.** These body part interaction pairs are chosen as discriminative (high positive weight w_i) for action classes indicated on the left. In each row, the first three images show detections on the correct action class. The last image shows a high scoring detection on an incorrect action class. In the examples shown, the interaction features capture either a body part and an object, or two body part interactions. Note that while these interaction pairs are found to be discriminative, due to the detection noise, they do not necessary localize the correct body parts in all images. However, they may still fire at consistent locations across many images as illustrated in the second row, where the head detector consistently detects the camera lens, and the thigh detector fires consistently at the edge of the head. Similarly, the leg detector seems to consistently fire on keyboards (see the third image in the first row for an example), thus improving the confidence of the computer detections for the "Interacting with computer" action.

Action / Method	d. Inter.	e. BOF+DPM+Inter.	Poselets	MK-KDA
(1) Phoning	42.1	48.6	49.6	**52.6**
(2) Playing Instr.	30.8	53.1	43.2	**53.5**
(3) Reading	28.7	28.6	27.7	**35.9**
(4) Riding Bike	**84.9**	80.1	83.7	81.0
(5) Riding Horse	89.6	**90.7**	89.4	89.3
(6) Running	81.3	85.8	85.6	**86.5**
(7) Taking Photo	26.9	**33.5**	31.0	32.8
(8) Using Comp.	52.3	56.1	59.1	**59.2**
(9) Walking	**70.1**	69.6	67.9	68.6
Average (mAP)	56.3	60.7	59.7	**62.2**

Table 4.2 – Per-class average-precision on the Pascal VOC 2010 action classification dataset. See text for description.

sification performance over a strong bag-of-features baseline in challenging realistic images of common human actions. In addition, the learned interaction features in some cases correspond to visually meaningful configurations of body parts, and body parts with objects.

We use only a small set of object detectors available at [Felzenszwalb et al., 2008a], however, we are now in a position to include many more additional object (camera, computer, laptop) or texture (grass, road, trees) detectors, trained from additional datasets, such as ImageNet or LabelMe. Currently, we consider detections of entire objects, but the proposed model can be easily extended to represent interactions between body parts and parts of objects [Brox et al., 2011].

In the next part of this thesis, we will focus on the opposite question: can we use persons to help object localization and scene understanding ? We will see that we can actually relate specific poses of people to specific actions and recover objects and layouts of indoor rooms.

Part II

Improving scene understanding

Chapter 5

Human actions and scene understanding

5.1 Introduction

In the previous chapter we used objects to improve human action recognition. Conversely, object functions can be derived from the known associations between object categories and human actions (the mediated perception of function approach [Palmer, 1999]), for example chair→sittable, window→openable. Actions such as sitting, however, can be realized in many different forms which can be characteristic for some objects but not for others, as illustrated in Figure 5.1. Moreover, some objects may not support the common function associated with their category: for example, windows in airplanes are usually not openable. These and numerous other examples suggest that the category-level association between objects and their functions is not likely to scale well to the very rich variety of the types and forms of person-object interactions. Instead, we argue that the functional descriptions of objects should be learned directly from observations of visual data.

In this chapter, we design object descriptions by learning associations between objects and spatially co-occurring human poses. To capture the rich variety of person-object inter-

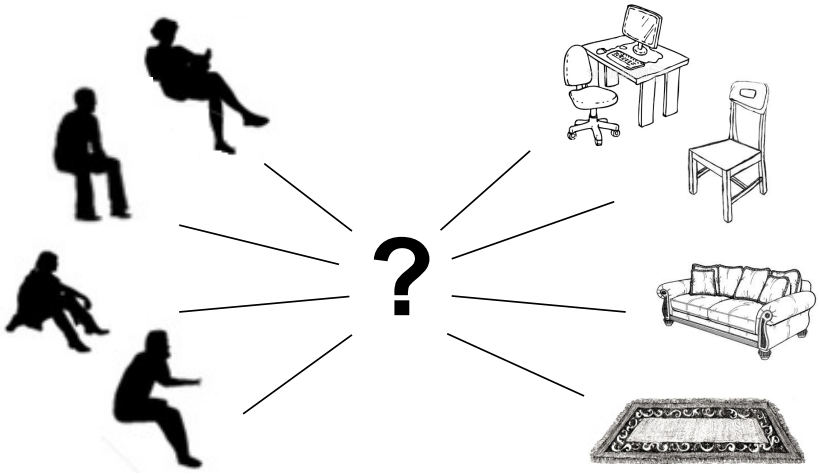


Figure 5.1 – Different ways of using objects. While all people depicted on the left are sitting, their sitting poses can be rather unambiguously associated with the objects on the right. In this chapter we build on this observation and learn object descriptions in terms of characteristic body poses.

actions, we automatically detect people and estimate body poses in long-term observations of realistic indoor scenes using the state-of-the-art method of [Yang and Ramanan, 2011] reviewed in Section 2.2. While reliable pose estimation is still a challenging problem, we circumvent the noise in pose estimation by observing many person interactions with the same instances of objects. For this purpose we use videos from hour-lasting events (parties, house cleaning) recorded with a static camera and summarized into time-lapses[1]. Static objects in time-lapses (e.g., sofas) can be readily associated with hundreds of co-occurring human poses spanning the typical interactions of people with these objects (see Figures 5.2-5.4). Equipped with this data, we construct statistical object descriptors which combine the signatures of object-specific body poses as well as the object's appearance. The model is learned discriminatively from many time-lapse videos of variety of scenes.

To summarize our contributions, we propose a new statistical model describing objects

[1]Time-lapse http://en.wikipedia.org/wiki/Time-lapse_photography is a common media type used to summarize recordings of long events into short video clips by temporal sub-sampling. We use time-lapses widely available on public video sharing web-sites such as YouTube, which are typically sampled at one frame per 1-60 seconds.

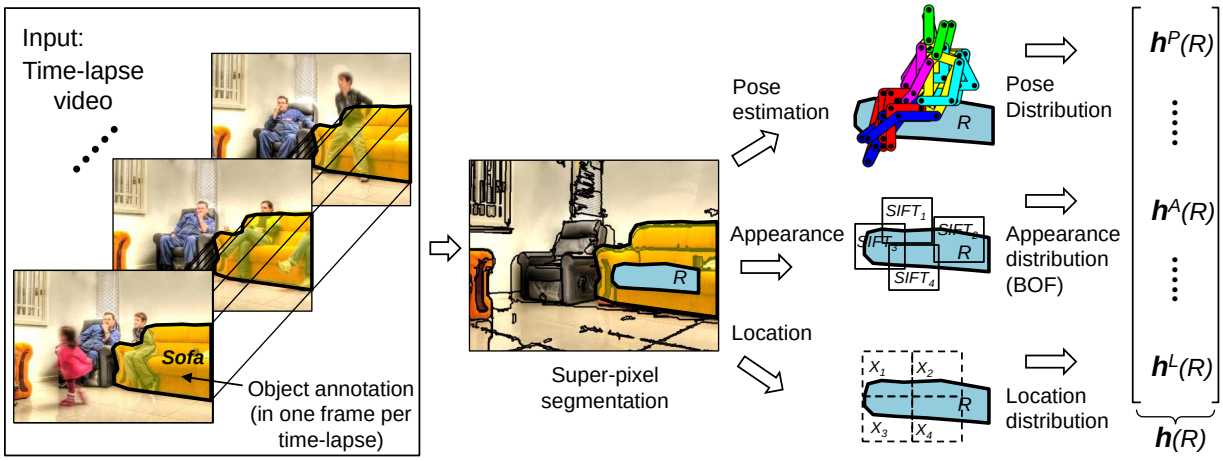


Figure 5.2 – Overview of the proposed person-based object description. Input scenes are over-segmented into super-pixels; each super-pixel (denoted R here) is described by the distribution of co-occurring human poses over time as well as by the appearance and location of the super-pixel in the image.

in terms of distributions of associated human poses. Notably, we do not require human poses to be annotated during training and learn the rich variety of person-object interactions automatically from long-term observations of people. Our functional object description generalizes across realistic and challenging scenes, provides significant improvements in object recognition and supports prediction of human poses in new scenes.

Chapter outline: We present an overview of the method in Section 5.2. We introduce our pose based descriptor in Section 5.3 and describe the appearance and location descriptors in Section 5.4. Section 5.5 focuses on learning the parameters of the model. Section 5.6 presents our *TimeLapse2D* dataset and Section 5.7 describes the experimental setup and the hyper-parameters we used to evaluate the object segmentation. It also presents qualitative results about the opposite task of predicting a pose given a the manually labeled object regions. Finally 5.8 sums up our contributions and present further directions of research.

5.2 Method overview

In this section we give a brief overview of the proposed approach. Our main goal is to learn functional object descriptions from realistic observations of person-object interactions. To simplify the learning task, we assume input videos contain static objects with fixed locations across their frames. Annotation of such objects in the whole video can be simply done by outlining object boundary in one video frame as illustrated in Figure 5.2. Moreover, person interactions with static objects can be automatically recorded by detecting people in the spatial proximity of annotated objects.

We start by over-segmenting input scenes into super-pixels, which will form the candidate object regions (details given in Section 5.5). For each object region R we construct a descriptor vector $\mathbf{h}(R)$ to be used for subsequent learning and recognition. The particular novelty of our method is a new descriptor representing an object region by the temporal statistics $\mathbf{h}^P(R)$ of co-occurring people (Section 5.3). This descriptor contains a distribution of human body poses and their relative location with respect to the object region. We also represent each object region by appearance features, denoted $\mathbf{h}^A(R)$, and the absolute location in the frame, denoted $\mathbf{h}^L(R)$, as described in Section 5.4.

Given descriptor vectors, one for each object region, containing statistics of characteristic poses, appearance and image locations, a linear support vector machine (SVM) classifier is learned for each object class from the labeled training data in a discriminative manner. At test time, the same functional and appearance representation is extracted from candidate object regions of the testing video. Individual candidate object regions are then classified as belonging to one of the semantic object classes.

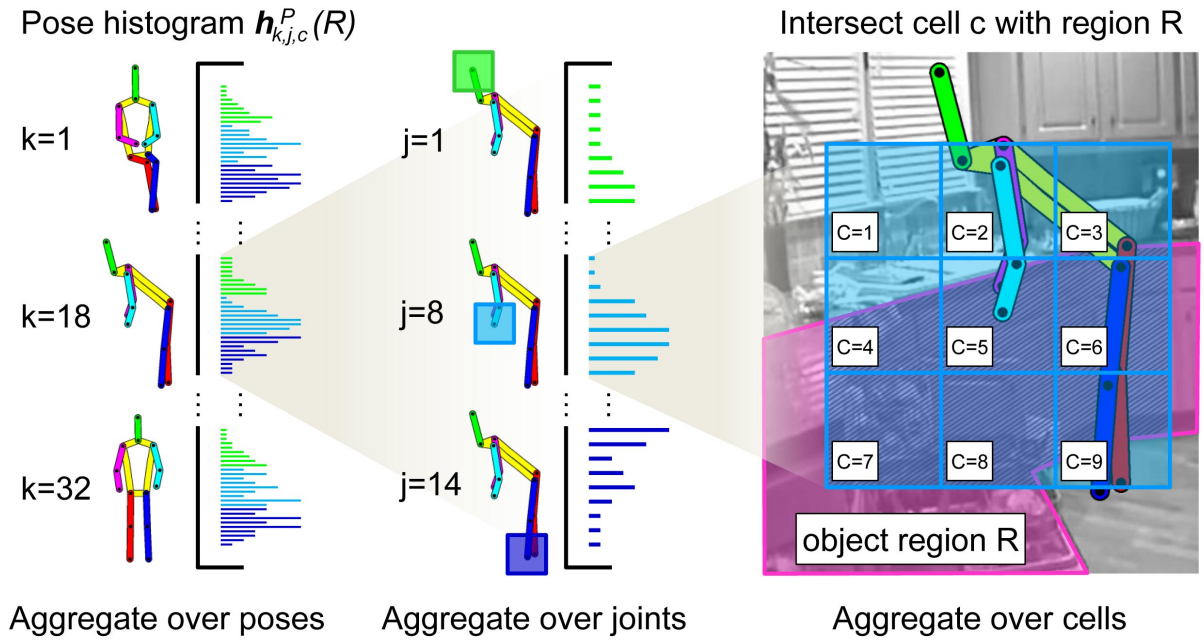


Figure 5.3 – Capturing person-object interactions. An object region R is described by a distribution (histogram) over poses k (left), joints j (middle) and cells c (right). The 3×3 grid of cells c is placed around each joint to capture the relative position of an object region R with respect to joint j. The pixel overlap between the grid cell c and the object region R weights the contribution of the j^{th} joint and the k^{th} pose cluster.

5.3 Modeling long-term person-object interactions

This section presents our model of the relationship between objects and surrounding people. We start by introducing a new representation describing an object by the statistics of co-occurring human poses. We then explain the details of the extraction and quantization of human poses in time-lapses.

5.3.1 Describing an object by a distribution of poses

We wish to characterize objects by the typical locations and poses of surrounding people. While 3D reasoning about people and scenes [Gupta et al., 2011] has some advantages, reliable estimation of scene geometry and human poses in 3D is still an open problem.

Moreover, deriving rich person-object co-occurrences from a single image is difficult due to the typically limited number of people in the scene and the noise of automatic human pose estimation. To circumvent these problems, we take advantage of the spatial co-occurrence of objects and people in the image plane. Moreover, we accumulate many human poses by observing scenes over an extended period of time.

In our setup we assume a static camera and consider larger objects such as sofas and tables which are less likely to change locations over time. We describe object region R in the image by the temporal statistics $\mathbf{h}^P$ of co-occurring human poses. Each person detection d is represented by the locations of $J(= 14)$ body joints, indexed by j, and the assignment q_k^d of d's pose to a vocabulary of K^P discrete pose clusters; see Figure 5.3 and Sections 5.3.2-5.3.3 for details. To measure the co-occurrence of people and objects, we define a spatial grid of 9 cells c around each body joint j. We measure the overlap between the object region R and the grid cell $B_{j,c}^d$ by the normalized area of their intersection $\mathcal{I}(B_{j,c}, R) = \frac{|B_{j,c} \cap R|}{|B_{j,c}|}$. We then accumulate overlaps from all person detections $\mathcal{D}$ in a given video and compute one entry $h_{k,j,c}^P(R)$ of the histogram descriptor $\mathbf{h}^P(R)$ for region R as

$$h_{k,j,c}^P(R) = \sum_{d \in \mathcal{D}} \frac{\mathcal{I}(B_{j,c}^d, R)}{1 + \exp(-3s_d)} \, q_k^d, \tag{5.1}$$

where k, j, and c index pose clusters, body joints and grid cells, respectively. The contribution of each person detection in (5.1) is weighted by the detection score s_d. The values of q_k^d indicate the similarity of the person detection d with a pose cluster k. In the case of the hard assignment of d to the pose cluster $\tilde{k}$, $q_k^d = 1$ for $k = \tilde{k}$ and $q_k^d = 0$ otherwise. In our experiments we found that better results can be obtained using soft pose assignment as described in the next section.

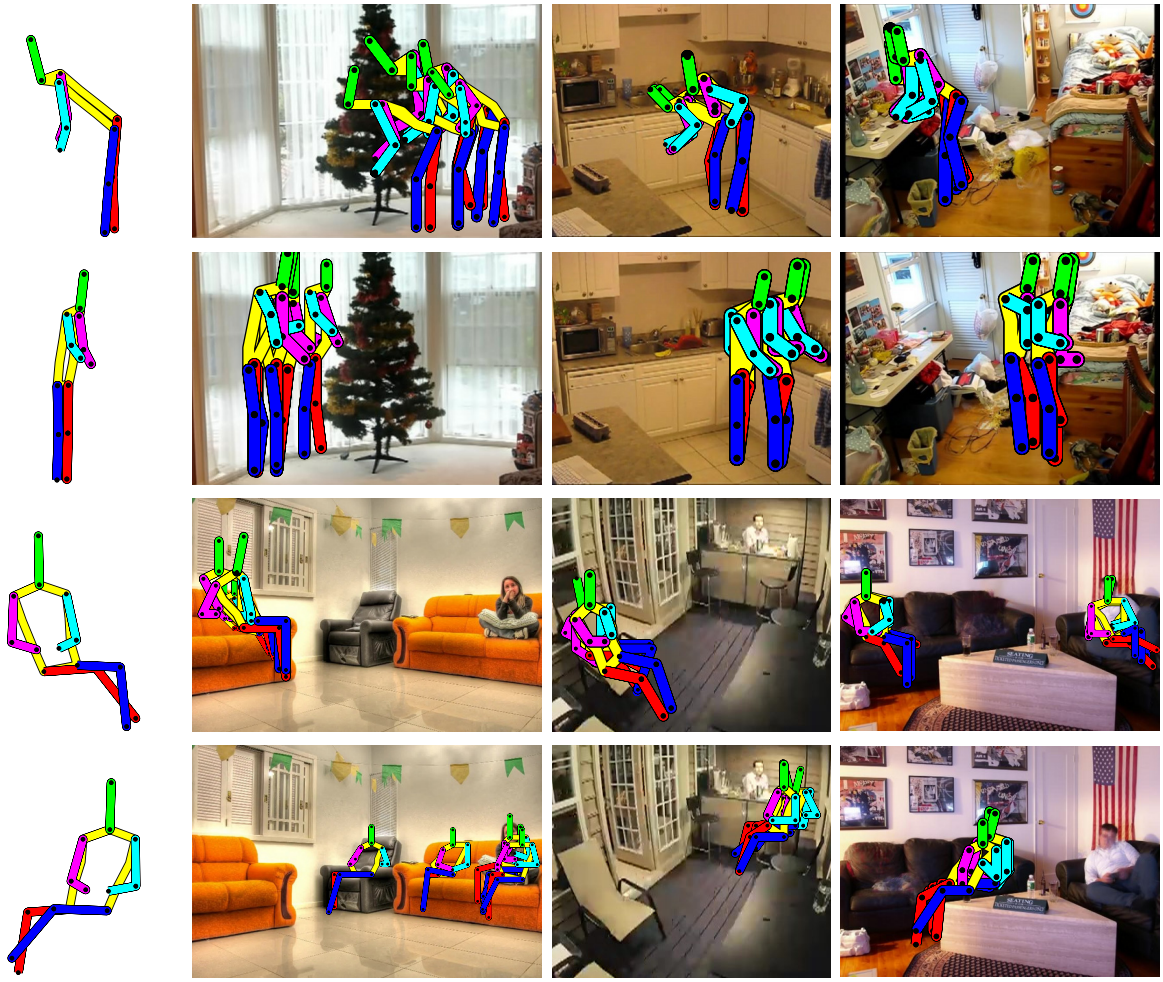


Figure 5.4 – Pose cluster and detection examples. Left: example cluster means from our pose vocabulary. Right: person detections in multiple frames of time-lapse videos assigned to the pose clusters on the left.

5.3.2 Building a vocabulary of poses

We represent object-specific human actions by a distribution of quantized human poses. To compute pose quantization, we build a vocabulary of poses from person detections in the training set by unsupervised clustering.

In order to build the pose vocabulary, we first convert each detection d in the training video into a $2J$-dimensional pose vector $\mathbf{x}^d$ by concatenating mid-point coordinates of all detected body joints. We center and normalize all pose vectors in the training videos

and cluster them by fitting a Gaussian Mixture Model (GMM) with K^P components via expectation maximization (EM). The components are initialized by the result of a K-means clustering and during fitting we constrain the covariances to be diagonal. The resulting mean vectors $\boldsymbol{\mu}_k$, diagonal covariance matrices $\boldsymbol{\Sigma}_k$ and weights π_k for each pose cluster $k = 1, \cdots, K^P$ form our vocabulary of poses (see Figure 5.4). A pose vector $\mathbf{x}^d$ for a detection d can be described by a soft assignment to each of the $\boldsymbol{\mu}_k$ by computing the posterior probability vector $\mathbf{q}^d$, where

$$q_k^d = \frac{p(\mathbf{x}^d|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k}{\sum_{j=1}^{K^P} p(\mathbf{x}^d|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}. \tag{5.2}$$

5.3.3 Person detection and pose estimation

We focus on detecting people in three body configurations common in indoor scenes: standing, sitting and reaching. We use the person detector from [Yang and Ramanan, 2011], which was shown to perform very well at both people detection and pose estimation and train three separate models, one for each body configuration. We found that training 3 separate models improved pose estimation performance over using a single generic pose estimator (Section 5.7).

The three detectors are run separately on all frames of each time-lapse video in a sliding window manner at multiple scales. As all our videos have fixed viewpoint, we use background subtraction (Section 5.7) to remove some false positive detections. Additional false positives can be removed via geometric filtering: we use the vanishing point estimation method proposed in [Hedau et al., 2009] to compute the horizon height y_h. We then assume a linear relationship $h_p(y_p) = \alpha(y_p - y_h)$ between a person's height h_p and the feet y-coordinate y_p in the image [Rodriguez et al., 2011], and learn the scaling coefficient α via RANSAC and robust least square fitting. We discard detections for which the difference

between the detected person height and the expected person height is greater than a given threshold ϵ. Finally we normalize the output of the detectors by making the mean and standard deviation of the detection scores equal to 0 and 1 on training videos, respectively. The filtering and normalization is performed separately for each detector.

To obtain the final set of detections, we perform standard non-maxima suppression on the combined outputs of the three detectors in each frame: if bounding boxes of several person detections overlap (i.e., have intersection over union bigger than 0.3), the detection with the highest normalized response is kept. This leads to a set $\mathcal{D}_i$ of confident person detections for the i^{th} video. Each detection $d \in \mathcal{D}_i$ is represented by an associated normalized score s^d and an estimated limb-configuration consisting of J bounding boxes B_j^d, $j = 1, \cdots, J$ corresponding to $J = 14$ locations of body joints.

As our time-lapse videos are sparsely sampled in time, the reasoning about temporal evolution of human poses is not straightforward. We therefore currently discard any temporal information about detected people. Nevertheless, the temporal re-occurrence of characteristic body poses for particular objects is a very powerful cue which we exploit to span the rich variety of person-object interactions.

5.4 Modeling appearance and location

In addition to the distribution of poses we also model the appearance and absolute position of image regions. Building on the work of [Sivic and Zisserman, 2003], we use an orderless bag-of-features representation and describe the appearance of image regions by a distribution of visual words. We first densely extract SIFT descriptors [Lowe, 2004] $f \in \mathcal{F}_k$ from image patches B^f of multiple sizes s_k for $k = 1, \cdots, S$ for all training videos and quantize them into visual words by fitting a GMM with K^A components. Each feature f is then soft-assigned to this vocabulary in the same manner as described in Eq. (5.2). This results

in an assignment vector $\mathbf{q}^f$ for each feature. The K^A-dimensional appearance histogram $\mathbf{h}^A(R)$ for region R is computed as a weighted sum of assignment vectors $\mathbf{q}^f$

$$\mathbf{h}^A(R) = \sum_{k=1}^{S} \sum_{f \in \mathcal{F}_k} s_k^2 \mathcal{I}(B^f, R)\, \mathbf{q}^f, \tag{5.3}$$

where $s_k^2 \mathcal{I}(B^f, R)$ is the number of pixels belonging to both object region R and feature patch B^f.

Similar to [Hoiem et al., 2005], we also represent the absolute position of regions R within the video frame. This is achieved by spatially discretizing the video into a grid of $m \times n$ cells, resulting in a $(m \times n)$-dimensional histogram $\mathbf{h}^L(R)$ for each region R. Here the i^{th} bin of $\mathbf{h}^L(R)$ is simply the proportion of pixels of the i^{th} cell of the grid falling into R.

5.5 Learning from long-term observations

We now detail how we obtain candidate object regions from multiple super-pixel segmentations and learn the model of person-object interactions. We then show how to recognize objects in testing videos and predict likely poses in new scenes.

5.5.1 Obtaining candidate object regions.

As described in previous sections, we represent objects by accumulating statistics of human poses, image appearance and location at object regions R. Candidate object regions are obtained by over-segmenting video frames into super-pixels using the method and on-line implementation of [Felzenszwalb and Huttenlocher, 2004]. As individual video frames may contain many people occluding the objects in the scene, we represent each video using a single "background frame" containing (almost) no people (Section 5.7). Rather than relying

on a single segmentation, we follow [Hoiem et al., 2005] and compute multiple overlapping segmentations by varying the parameters of the segmentation algorithm.

5.5.2 Learning object model.

We train a classifier for each object class in a one-versus-all manner. The training data for each classifier is obtained by collecting all (potentially overlapping) super-pixels, R_i for $i = 1, \cdots, N$, from all training videos. For each region, we extract their corresponding pose, appearance and location histograms as described in Sections 5.3 and 5.4. The histograms are separately L_1-normalized and concatenated into a single K-dimensional feature vector $\mathbf{x}_i = [\tilde{\mathbf{h}}^P(R_i), \tilde{\mathbf{h}}^A(R_i), \tilde{\mathbf{h}}^L(R_i)]$, where $\tilde{\mathbf{h}}$ denotes L_1-normalized histogram $\mathbf{h}$. An object label y_i is then assigned to each super-pixel based on the surface overlap with the provided ground truth object segmentation in the training videos. Using the surface overlap threshold of 34%, each super-pixel can be assigned up to two ground truth object labels. Finally we train a binary support vector machine (SVM) classifier with the Hellinger kernel for each object class using the labeled super-pixels as training data. The Hellinger kernel is efficiently implemented using the explicit feature map $\Phi(\mathbf{x}_i) = \sqrt{\mathbf{x}_i / L_1(\mathbf{x}_i)}$ and a linear classifier. Finally, the outputs of individual SVM classifiers are calibrated with respect to each other by fitting a multinomial regression model from the classifiers output to the super-pixel labels [Hastie et al., 2003]. The output of the learning stage is a K-dimensional weight vector $\mathbf{w}_y$ of the (calibrated) linear classifier for each object class y.

At test time, multiple super-pixel segmentations are extracted from the background frame of the test video and the individual classifiers are applied to each super-pixel. This leads to a confidence measure for each label and super-pixel. The confidence of a single image pixel is then the mean of the confidences of all the super-pixels it belongs to.

5.5.3 Inferring probable pose.

Here we wish to predict the most likely pose within a manually provided bounding box in an image, given an object layout (segmentation) of the scene. This is achieved by choosing the pose cluster, for which the sum of learned object weights for all joints most agree with the given per-pixel object labels in the image. More formally, denoting $w_y(k, j, c)$ the weight learned for label y, pose cluster k, joint j and grid cell c, we select the pose cluster $\hat{k}$ that maximizes the sum of per-pixel weights under each joint grid cell $B^k_{j,c}$

$$\hat{k} = \arg\max_k \sum_{j=1}^{J} \sum_{c=1}^{9} \sum_{\text{pixels } i \in B^k_{j,c}} w_{y_i}(k, j, c), \tag{5.4}$$

where y_i is the label for pixel i.

5.6 The TimeLapse2D dataset

We extend the dataset of [Fouhey et al., 2012] to 146 time-lapse videos containing a total of around 400,000 frames. Each video sequence shows human actors interacting with an indoor scene over a period of time ranging from a few minutes to several hours. The captured events include parties, working in an office, cooking or room-cleaning. The videos were downloaded from YouTube by placing queries such as "time-lapse party". Search results were manually verified to contain only videos captured with a stationary camera and showing an indoor scene. All videos are sparsely sampled in time with limited temporal continuity between consecutive frames. The dataset represents a challenging uncontrolled setup, where people perform natural non-staged interactions with objects in a variety of real indoor scenes.

We manually annotated each video with ground truth segmentation masks of eight frequently occurring semantic object classes: *'Bed'*, *'Sofa/Armchair'*, *'Coffee Table'*, *'Chair'*,

'Table', *'Wardrobe/Cupboard'*, *'Christmas tree'* and *'Other'*. Similar to [Hedau et al., 2009], the *'Other'* class contains various foreground room clutter such as clothes on the floor, or objects (e.g., lamps, bottles, or dishes) on tables. In addition to objects we also annotated three room background classes: *'Wall'*, *'Ceiling'* and *'Floor'*. As the camera and majority of the objects are static, we can collect hundreds or even thousands of realistic person-object interactions throughout the whole time-lapse sequence by providing a single object annotation per video. The dataset is divided into 5 splits of around 30 videos with approximately the same proportion of labels for different objects. The TimeLapse2D dataset including the annotations is available at http://www.di.ens.fr/willow/research/scenesemantics/.

5.7 Experiments

In this section we give the implementation details and then show results for (i) pose estimation (ii) semantic labeling of objects in time-lapse videos and (iii) predicting likely poses for new scenes.

Implementation details.

The foreground/background segmentation in each video frame is estimated using a pixel-wise adaptive mixture of Gaussian with 5 components [Staufer and Grimson, 1998] (with $\alpha = 0.01$ and $T = 0.2$). We also compute a single "background image" for each video that contains no people by taking the median of background segments across all video frames. Person detections and human pose estimates in each frame are obtained using the method and code of [Yang and Ramanan, 2011]. Detections in the background segments and with confidence smaller than -1.1 are removed. The threshold ϵ for the ground-plane based geometric filter [Rodriguez et al., 2011] is set to 30%. Super-pixels for each video are generated using the code of [Felzenszwalb and Huttenlocher, 2004] with parameters

$\sigma \in \{0.2, 0.3\}$, $k = 80$ and $min = 600$. SIFT features are extracted from patches of size $s \in \{8, 16, 32, 64\}$ pixels, with 50% spatial overlap. To train the proposed model, we use 3 splits of the dataset (see Section 5.6) to cross-validate the C parameter of the SVM and use the 4^{th} split to calibrate the outputs of the individual classifiers. The resulting model is tested on the 5^{th} split. This is repeated five times for the different test splits to obtain the mean and standard deviation of the classification performance.

Pose estimation.

To evaluate person detection and pose estimation performance we have annotated poses of at least ten (randomly chosen) person occurrences in each video, resulting in 1606 pose annotations. Person (bounding box) detection performance is measured using the standard average precision (AP) and pose estimation performance is measured by the Percentage of Correct Parts (PCP) score among the detected people as proposed in [Ferrari et al., 2008b]. We first compare our individually trained pose estimators for each action (see Section 5.3.3) with a single model trained on images from all 3 action classes. Both have a similar recall of around 52% but the individually trained models achieve an average PCP of 50% compared to 47% for the single model. We then evaluate the effect of the background subtraction and geometric filtering for person detection. The individually trained models achieve an AP of 33%, which is significantly improved by background subtraction (51%) and geometric filtering (56%).

Semantic labeling of objects.

Semantic labeling performance is measured by pixel-wise precision-recall curve and average precision (AP) for each object. Table (5.1) shows the average precision for different object and room background classes for different feature combinations of our method. Performance is compared to two baselines: the method of [Hedau et al., 2009], trained

	DPM	Hedau	(A+L)	(P)	(A+P)	(A+L+P)
Bed	**31±20**	12±7.2	14±5.0	21±5.8	27±13	26±13
Sofa/Armchair	26±9.4	26±10	34±3.3	32±6.5	**44±5.4**	43±5.8
Coffee Table	11±5.4	11±5.2	11±4.4	12±4.3	**17±10**	**17±9.6**
Chair	9.5±3.9	6.3±2.8	8.3±2.7	5.8±1.4	11±5.4	**12±5.9**
Table	15±6.4	18±3.8	17±3.9	16±7.1	**22±6.2**	**22±6.4**
Wardrobe/Cupboard	27±10	27±8.2	28±6.4	22±1.1	**36±7.4**	**36±7.2**
Christmas tree	50±3.3	55±12	72±1.8	20±6.0	76±6.2	**77±5.5**
Other Object	12±6.4	11±1.2	7.9±1.9	13±4.2	**16±8.3**	**16±8.2**
Average (objects only)	23±1.8	21±1.5	24±2.0	18±2.2	**31±4.6**	**31±4.8**
Wall	—	75±3.9	76±1.6	76±1.7	**82±1.2**	81±1.3
Ceiling	—	47±20	53±8.0	52±7.4	**69±6.7**	**69±6.6**
Floor	—	59±3.1	64±5.5	65±3.6	**76±3.2**	**76±2.9**
Average (all classes)	—	31±2.0	35±2.4	30±1.7	**43±4.4**	**43±4.3**

Table 5.1 – Average precision (AP) for baselines of [Felzenszwalb et al., 2009] and [Hedau et al., 2009] compared to four different settings of our method: appearance and location features only (A+L), person features only (P), appearance and person features (A+P), appearance, location and person features combined (A+L+P).

on our data with semantic object annotations, and the deformable part model (DPM) of [Felzenszwalb et al., 2009] trained over manually defined bounding boxes for each class. At test time, the DPM bounding boxes are converted to segmentation masks by assigning to each testing pixel the maximum score of any overlapping detection. Note that combining the proposed pose features with appearance (A+P) results in a significant improvement in overall performance, but further adding location features (A+L+P) brings little additional benefit, which suggests that spatial information in the scene is largely captured by the spatial relation to the human pose. The proposed method (A+L+P) also significantly outperforms both baselines. Example classification results for the proposed method are shown in Figure 5.5. Finally, learned weights for different objects are visualized in Figure 5.6.

We have also evaluated our model on functional surface estimation. For training and testing, we have provided ground truth functional surface masks for the dataset of [Fouhey et al., 2012]. Our model achieves AP of 76%, 25% and 44% for *'Walkable'*,

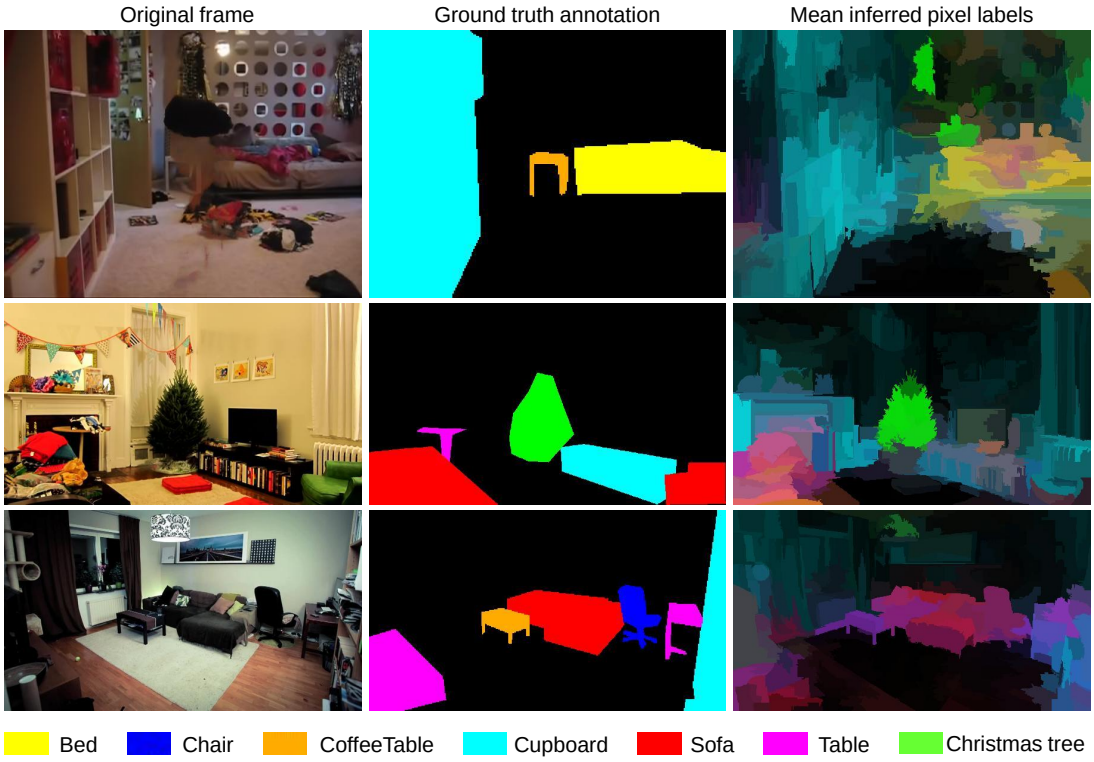


Figure 5.5 – Object soft segmentation. Scene background with no people (left). Object ground truth (middle). Mean probability map for inferred objects (right).

'Sittable' and *'Reachable'* surfaces, respectively, averaging a gain of 13% compared to Fouhey et al., which could be attributed to the discriminative nature of our model.

Predicting poses in new scenes.

Figure 5.7 shows qualitative results of predicting likely human poses in new scenes. Given a person bounding box and the manually labeled object regions, the most likely pose is predicted using Equation (5.4). As can be seen, the automatically generated poses are consistent with object classes as well as with the scene geometry despite no explicit 3D reasoning is included in our model.

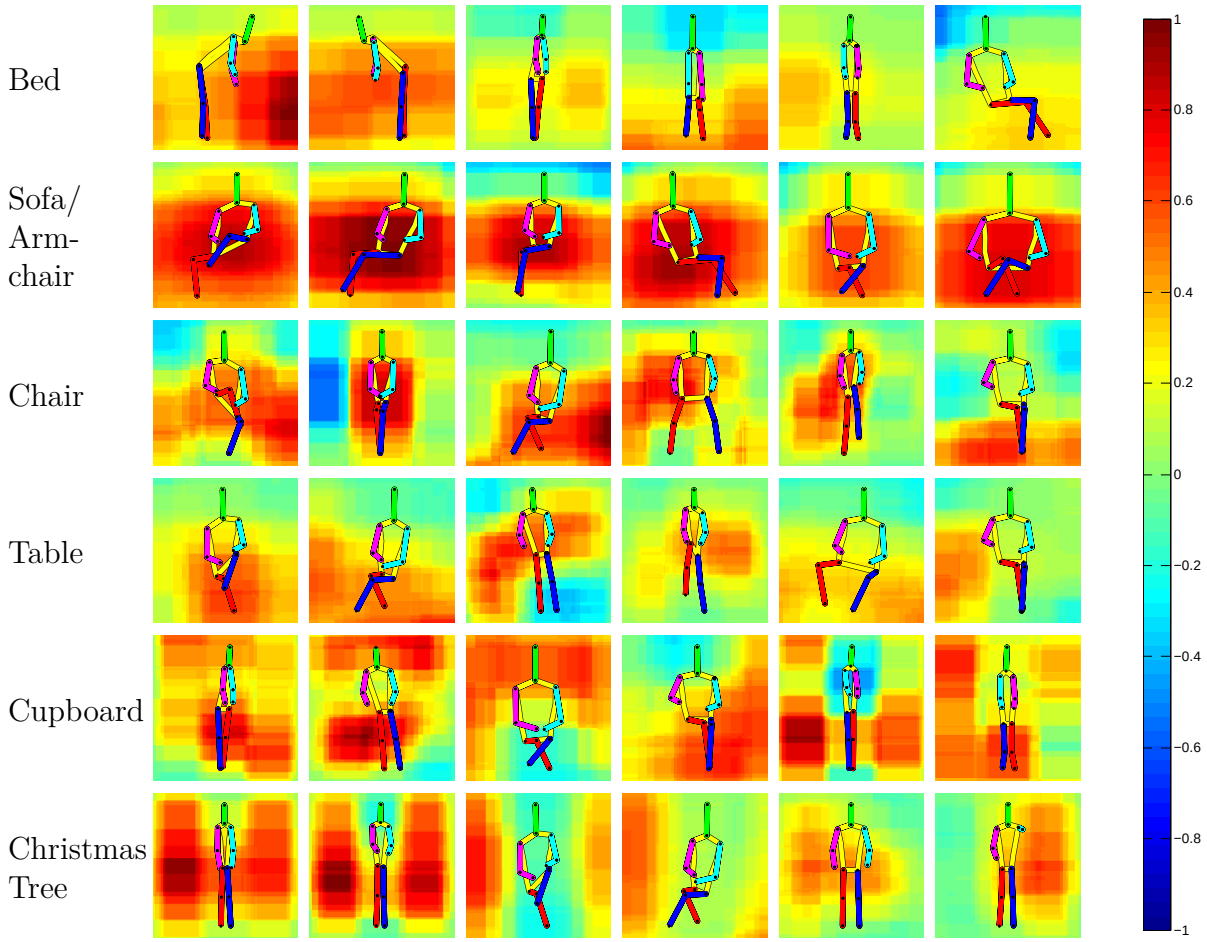


Figure 5.6 – Spatial locations of objects relative to particular poses. The top 6 pose clusters with the highest sum of positive weights are shown for selected objects (rows). Color indicates the spatial weights for the position of a given object relative to the particular body pose summed over all 9 grid cells for all joints. The color map is shown on the right. Note how, for example, Sofa/Armchair is likely to be located behind sitting people (2nd row) and table in the vicinity of sitting and standing people (4th row). The top scoring sitting poses for Sofa/Armchair are also quite different (more relaxed) than the top scoring sitting poses for Chair.

5.8 Discussion

We have proposed a statistical descriptor of person-object interactions and have demonstrated its benefits for recognizing objects and predicting human body poses in new scenes. Notably, our method requires little annotation and relies on long-term observations of people in time-lapse videos. In the next chapter, we use this model combined with room layout estimation to estimate the 3D volumes occupied by different object classes in indoor scenes.

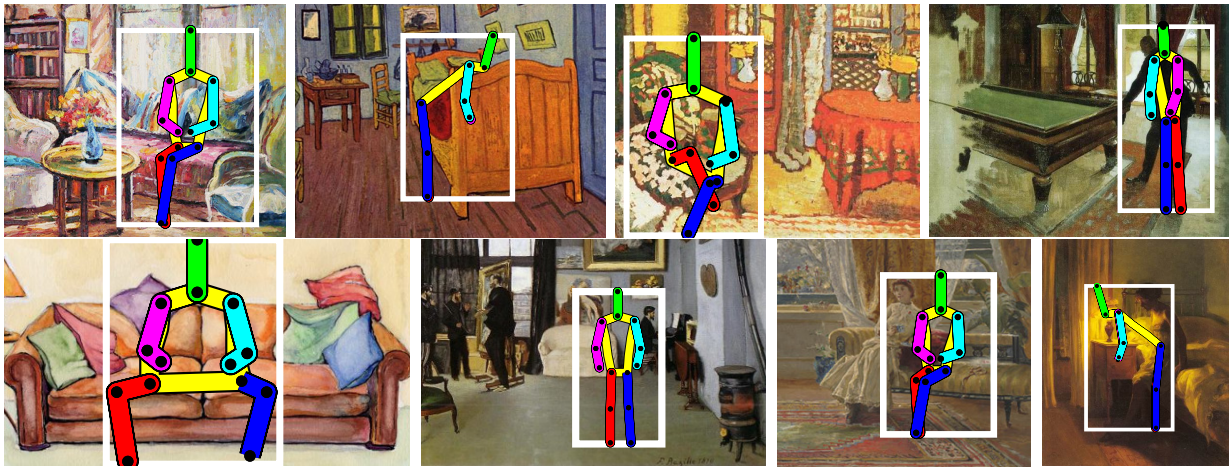


Figure 5.7 – Plausible poses prediction. The proposed model supports automatic prediction of plausible human poses in new scenes. This is achieved by selecting a pose cluster leading to the best agreement between the (manually provided) scene object layout and the object weights learned for each joint.

Chapter 6

People and semantic 3D geometry estimation

6.1 Introduction

In the previous chapter, we have developed a method for improving the object recognition in indoor environments based on the observation of people. We now focus on the wider problem of indoor scene understanding. We first describe a challenging dataset for evaluating the different steps of current methods for scene understanding. We then investigate how person cues can help inferring the room layout and semantic 3D space occupancy, i.e. the object labels of occupied 3D space volumes in a room.

Most of the current pipelines for indoor scene understanding reviewed in Section 2.4 consist of the following steps:

S1) **Vanishing point estimation**, see Figure 6.1(a): the goal of this step is to estimate the intrinsic camera parameters. Under the hypothesis that the directions of lines in an indoor scene image mainly correspond to the three orthogonal directions of the

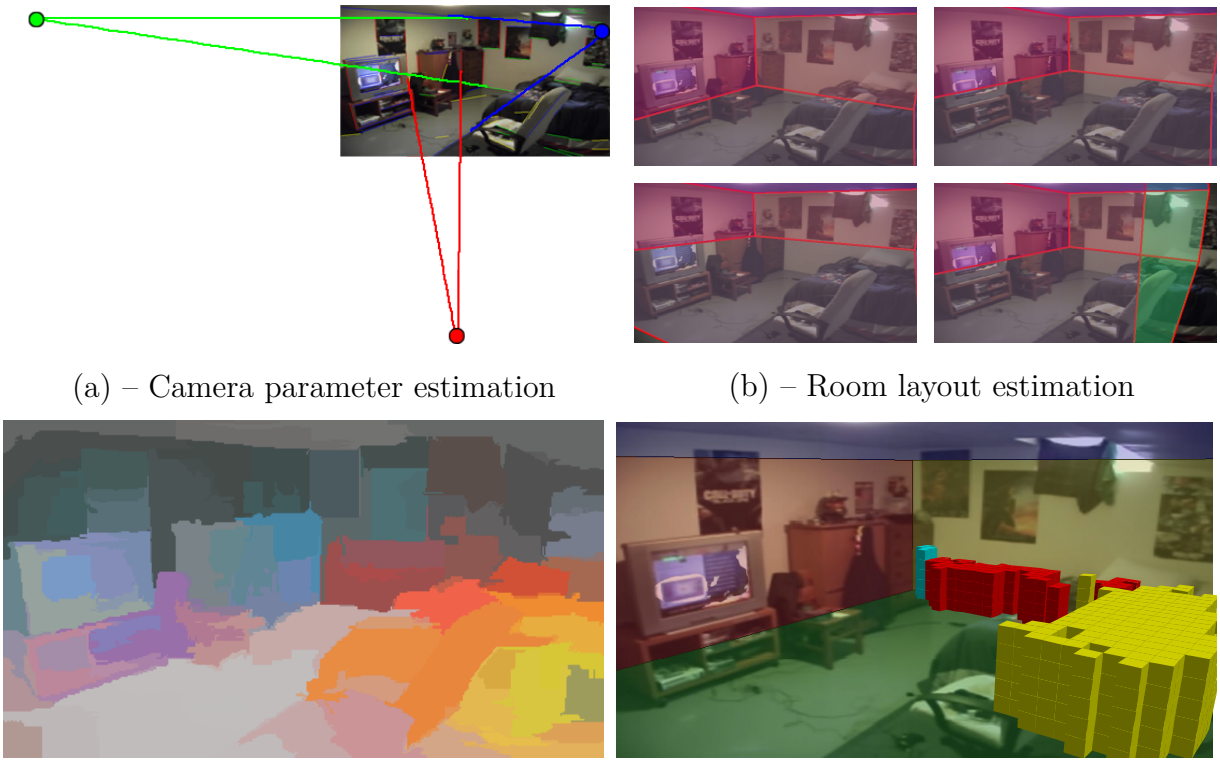
room, one can recover the corresponding vanishing point coordinates and the focal length of the camera, assuming the camera has square pixels and zero skew. The errors for this step typically come from an incorrect estimation of the three dominant directions (typically the case for cluttered rooms) or from the fact that the problem may be under-contrained when the vanishing points are located at the infinity.

S2) **Room layout estimation**, see Figure 6.1(b): Assuming that rooms are boxes aligned with the 3 previously computed vanishing points, the goal of this step is to generate several room layouts (i.e. wall positions in 3D) and select the best scoring one according to image cues. Problems often arise due to errors in the previous step of vanishing point estimation or because of the too coarse room layout sampling.

S3) **2D Object segmentation**, see Figure 6.1(c): This step aims at assigning each pixel (or super-pixel) of the image to an object class or to background. Typical errors are due to pixel misclassification or to super-pixels leaking outside of object boundaries.

S4) **3D space occupancy estimation**, see Figure 6.1(d): By combining the two previous steps, the goal of this part is to provide a finer understanding of the room in 3D by estimating the occupied voxels. Apart from errors due to the previous steps, this part may give inaccurate results due to geometric simplifications in the model, e.g. due to the quantization of the space into voxels or to the assumption that all objects of a given class share the same average height, see Subsection 6.3.4.

The measurement of the errors accumulated by each of the steps S1-S4 above has not been addressed until now probably because most of the experiments conducted so far used rooms with limited amount of clutter and occlusions. For example, [Choi et al., 2013] report promising accuracy on their dataset, however application of their method to our TimeLapse3D dataset containing more realistic and more complex time-lapse videos resulted in a significant drop in performance (see images of rooms from both datasets in



(a) – Camera parameter estimation (b) – Room layout estimation

(c) – 2D Object localization (d) – 3D Object localization

Figure 6.1 – Consecutive steps of the standard room understanding pipeline: (a) Camera calibration: estimated directions (lines) and vanishing points (dots) for the 3 orthogonal directions (red, green, blue), (b) Room layout estimation: four room layouts of different sizes, best layout is top-left, (c) Example results for 2D object soft-segmentation, high color intensity indicates high confidence: yellow for bed, red for sofa, cyan for cupboard, magenta for table, orange for coffee table, gray scale for floor, walls and ceiling, (d) 3D object localization: visualization of occupied voxels, same color code as (c). Please refer to text for details.

Figure 6.2). Similarly, [Fouhey et al., 2014] argue that we may be missing reliable 3D primitives for robust 3D scene understanding in cluttered rooms. A related problem comes from the fact that the performance measure generally used, i.e. the wall layout and "clutter" estimation accuracy measured pixel-wise in the image plane, may not reflect the actual 3D fit between the proposed layout and the ground truth room [Hedau et al., 2012] due to the scale ambiguity inherent to monocular room understanding. Most of the current work on 3D space occupancy is restricted to the estimation of "3D clutter". Here we argue for

the importance of classification of occupied space to corresponding object categories and propose an evaluation scheme for this tasks.

While previous datasets [Hedau et al., 2009, Hedau et al., 2010, Choi et al., 2013] contain scenes with no people, our more realistic dataset contains scenes with people which typically present challenges for common methods. People, however provide additional cues which can be explored for scene understanding. For example in Figure 6.3, most people can easily tell that the two depicted poses originate from room **A**. The pose of the left figure indeed reveals a horizontal surface right under its pelvis ending abruptly at its knees. The pose of the right figure reveals a ground plane under its feet as well as a likely horizontal surface near the hand location. The scale of both people also imposes a constraint on the size of nearby objects. In this chapter, we want to take advantage of the strong physical and functional coupling between people and the geometry of the scene to improve the different steps of the room understanding pipeline S1-S4 above.

We therefore propose an evaluation of errors in standard pipelines for 3D room understanding and use our TimeLapse3D dataset of cluttered scenes, based on the 2D pixel accuracy measure and two additional 3D measures. Building on the previous chapter, we evaluate the influence of person cues on the performance of room estimation and evaluate a new task of 3D semantic object localization, referred to as "3D semantic space occupancy estimation". To our knowledge, this is the first attempt to address this problem.

For the purpose of evaluation, we use the challenging dataset presented in the previous chapter with extended annotations. We manually annotate vanishing points, camera calibration, 3D room layout and 3D object positions. This allows us to measure the performance of each step of the method, from camera calibration to 3D voxel occupancy, for different experimental setups.

To sum up, the contributions of this chapter are three-fold: (a) we collect 3D semantic annotations for our time-lapse dataset in order to evaluate *semantic* 3D space occupancy

Figure 6.2 – Youtube consumer time-lapse videos tend to be more cluttered and have more occlusions than the clean images usually present in the standard datasets for 3D indoor scene understanding. Top row: the dataset of [Choi et al., 2013], bottom row: our dataset, see Section 6.2. For both rows, from left to right: a bedroom, a dining room and a living room.

estimation, (b) we separately evaluate the importance of each step S1-S4 on the final performance for 3D space occupancy estimation, (c) we evaluate how person cues can help different parts of this pipeline. Comparing to our previous work [Fouhey et al., 2012] where we were not evaluating the results of space occupancy estimation, we address in this chapter the evaluation of how well we can estimate the semantic labels of 3D volumes.

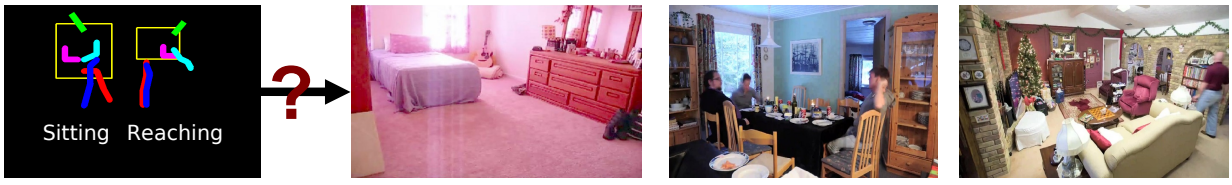


Figure 6.3 – Human actions tell us a lot about the 3D structure of a scene. Image on the left shows real detections taken from one of the 3 scenes of the right. The scale and pose of those detections impose constraints on the layout and objects in the scene (see text for more detail). These constraints are only satisfied by scene **A**.

Chapter outline: The rest of this chapter is organized as follows: Section 6.2 presents

the augmented 3D time-lapse dataset. In Section 6.3 we detail each step of the standard pipeline for 3D space estimation. Section 6.4 introduces performance measures. We present and discuss results of experiments in Section 6.5 and conclude this chapter in Section 6.6.

6.2 The TimeLapse3D dataset

We have enriched the TimeLapse2D dataset of the previous chapter with semantic 3D annotations. This dataset contains 146 time-lapse videos of indoor scenes with static cameras covering a long period of time in only few minutes by sparsely sub-sampling video frames. Each sequence shows people interacting with objects in the room while e.g. sitting, cleaning or partying.

To extend the TimeLapse2D dataset with 3D annotations, we have designed an annotation tool[1] that allows a user to drag and drop 3D object models in a 3D box of a room. First, a calibration step is performed. We ask the user to click on the corners of the room in order to compute the ground truth vanishing points and room dimensions. Next, the user is asked to position and resize a 3D object so that its projection in the camera plane fits the annotated image, see Figure 6.4 for a screenshot of the tool. We make the hypothesis that objects are axis-aligned, which is true for most of the rooms in the dataset. Most of the previous works have arbitrary fixed the camera height to solve the scale ambiguity, e.g. [Fouhey et al., 2012, Satkin et al., 2012, Hedau et al., 2009, Hedau et al., 2010, Hedau et al., 2012]. In our case the camera height is highly variable, e.g. some cameras are attached to the ceiling. We instead fix the room height to be 2.7 meters, except for three videos having a high ceiling that we fix at 3.5 meters.

Using our 3D annotation tool, we have annotated 146 videos of the dataset with ground truth vanishing points, 3D room walls and 3D objects. We have extended the 8 object cat-

[1] Available at https://github.com/vdel/RoomAnnotTool.

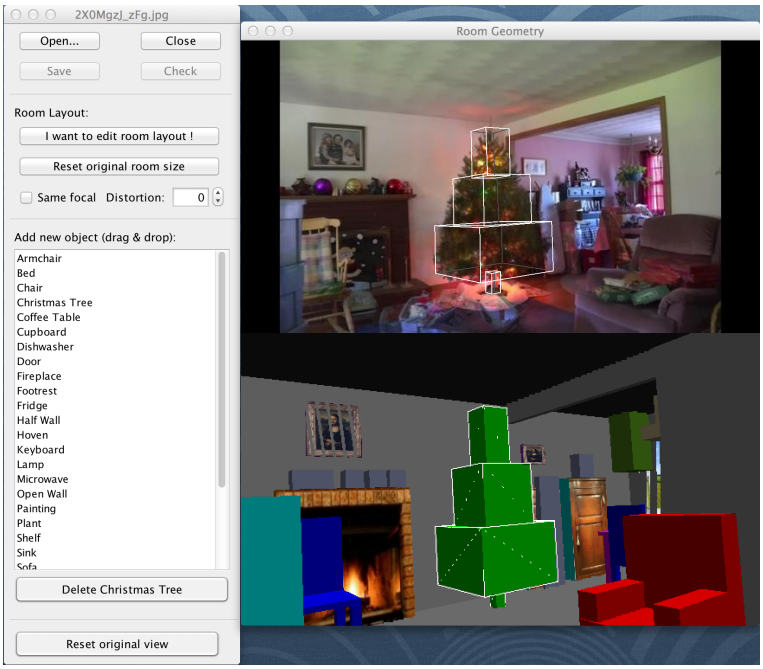


Figure 6.4 – Screenshot of our 3D annotation tool. After clicking on the corners of the room, the user can drag objects from the left panel and drop them in the virtual 3D scene (bottom half of the right panel). He can then place, rotate, and resize the object so that their projection in the image plane (top half of the right panel) fits the scene picture. In this example we selected the Christmas tree.

egories of the original time-lapse dataset to 26 object classes[2]. In the following evaluation, however, we restrict ourselves to the same classes used in the previous chapter, namely *'Bed'*, *'Sofa/Armchair'*, *'Coffee Table'*, *'Chair'*, *'Table'*, *'Wardrobe/Cupboard'*, *'Christmas tree'* and *'Other object'*. Examples of annotated rooms are depicted in Figure 6.5. This *TimeLapse3D* dataset is made available at https://github.com/vdel/TimeLapse3D.

[2]Our dataset includes the following labels: *Bed, Sofa, Armchair, Coffee Table, Chair, Foot rest, Table, Wardrobe/Cupboard, Fireplace, TV, Painting/Poster, Lamp, Suspended Lamp, Window, Door, Fridge, Hoven, Microwave, Dishwasher, Shelf, Keyboard, Washing Machine, Sink, Plant, Christmas tree, Other object*

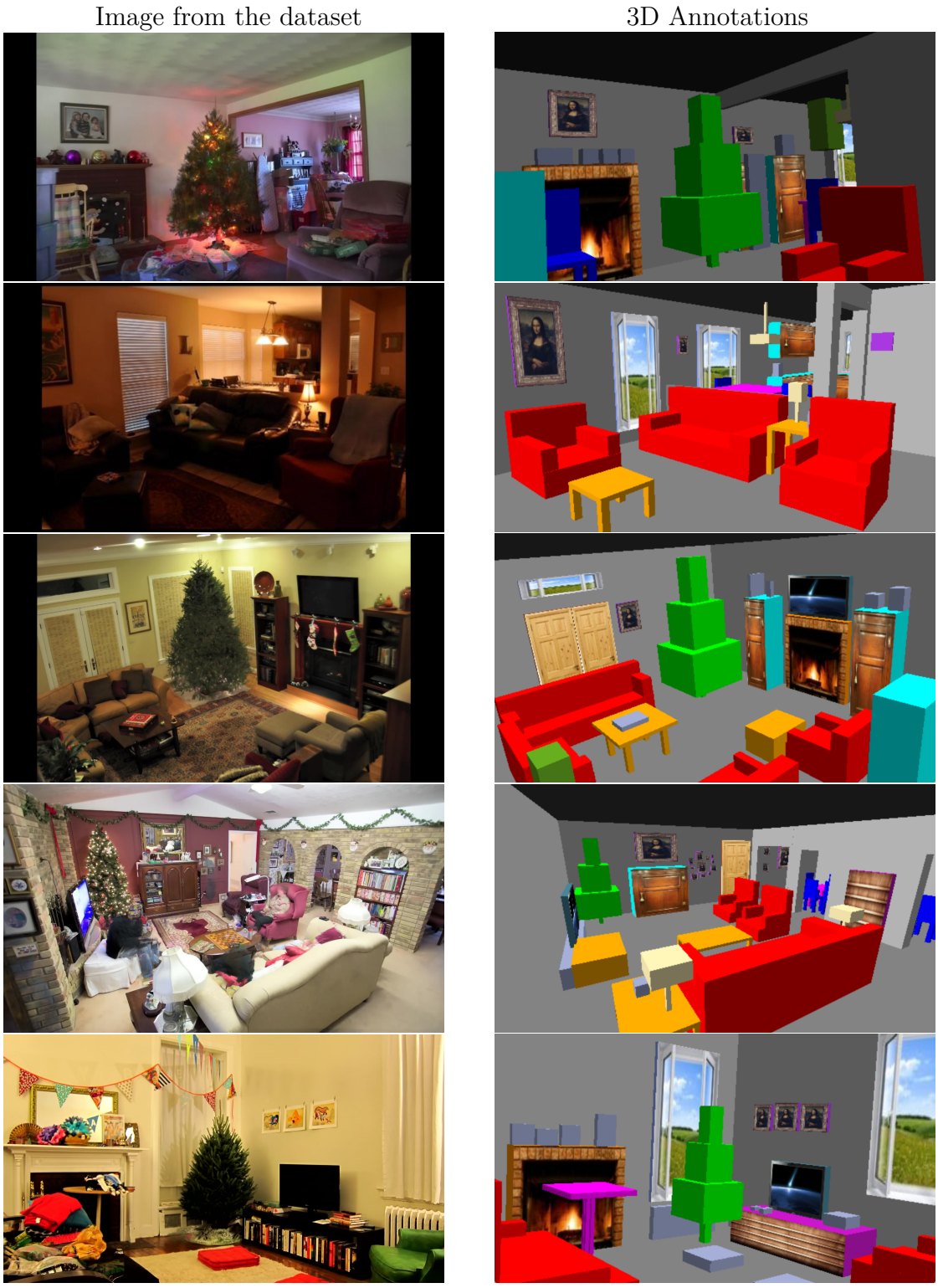


Figure 6.5 – Our 3D annotated time-lapse dataset contains a variety of rooms and objects, e.g. sofas and armchairs in red, coffee tables and foot rests in orange, tables in magenta, cupboards in cyan and chairs in navy-blue.

6.3 Estimation of semantics in 3D

We aim to estimate 3D object volumes and their corresponding class labels. We refer to this task as *3D semantic space occupancy*. We build on the previous work to: (a) measure the impact of each of the four steps S1-S4 on the 3D semantic space occupancy estimation (see Section 6.1) and (b) investigate the benefits of using person cues to improve those steps. More precisely, we use the algorithm of [Hedau et al., 2009] for steps S1 and S2, combined with the methods described in our previous work [Fouhey et al., 2012] (but not presented in this thesis) for steps S2 and S4. We use the method described in the previous chapter for step S3. In the following we detail each of those steps.

6.3.1 Camera calibration (S1)

Following the approach of [Hedau et al., 2009], we first extract lines in the image and group them in 3 main orthogonal directions, see Figure 6.6(a). We compute the associated vanishing points by letting the lines of each group vote following an exponential voting scheme [Hedau et al., 2009], see Figure 6.6(b). We then compute the camera calibration following the method of [Hedau et al., 2009]. Assuming that the camera has zero skew and square pixels, we estimate the focal length so that the axis of the room are orthogonal, which gives the projection matrix K. We deduce the rotation matrix R from the vanishing points.

The above procedure for camera calibration may fail due to distortions of the image (e.g. radial distortion) and invalid assumptions of "Manhattan" world geometry. In those cases, using people for camera calibration is not expected to help. Indeed, in presence of high distortion there is no linear relation between the 3D homogeneous position and image position of objects. For scenes which break the Manhattan world assumption one could rely on people to estimate the horizon line but the vanishing points for horizontal

directions still cannot be uniquely defined. For those reasons, we do not investigate the impact of people on step S1 but only the impact of using the ground truth calibration on the consecutive steps.

6.3.2 Scene layout selection and re-ranking (S2)

During this step, we aim at generating the room layout hypotheses. Following the method of [Hedau et al., 2009], we assume that at most three walls are visible and we define a room layout by the positions of the floor, the ceiling and these three walls. As we know the positions of the vanishing points corresponding to the directions of the walls, it is sufficient to know the extent of the most central wall to deduce the whole layout. This central wall can in turn be uniquely defined by its sides, thus by two rays originating from the vanishing point corresponding to vertical directions and two rays originating from the vanishing point corresponding to horizontal directions. In order to generate many layout hypotheses, we sample a set of rays originating from those two vanishing points and covering the whole span of the image, see Figure 6.6(c). Comparing to the dataset of [Hedau et al., 2009], our rooms tend to be deeper and we had to reduce by two the angle between two consecutive rays (see Figure 6.7 for an illustration). We pick different combinations of four rays to produce the layout hypotheses, see Figure 6.6(d). Given the image positions of the front wall's corners, we then compute the 3D position of the camera's center c_i (used in step S4) with respect to room origin for each generated room hypothesis. To solve for the scale ambiguity, we fix the room height to be 2.7 meters rather than fixing the camera height as done in other works reviewed in Section 2.4. We took this decision because the camera height is very variable in our dataset. The most promising layout is selected based on visual features as done in [Hedau et al., 2009], see Figure 6.6(e).

One can also use people to re-score different layouts as we did in [Fouhey et al., 2012]. This method uses the detections of people in an image or time-lapse and casts votes at the

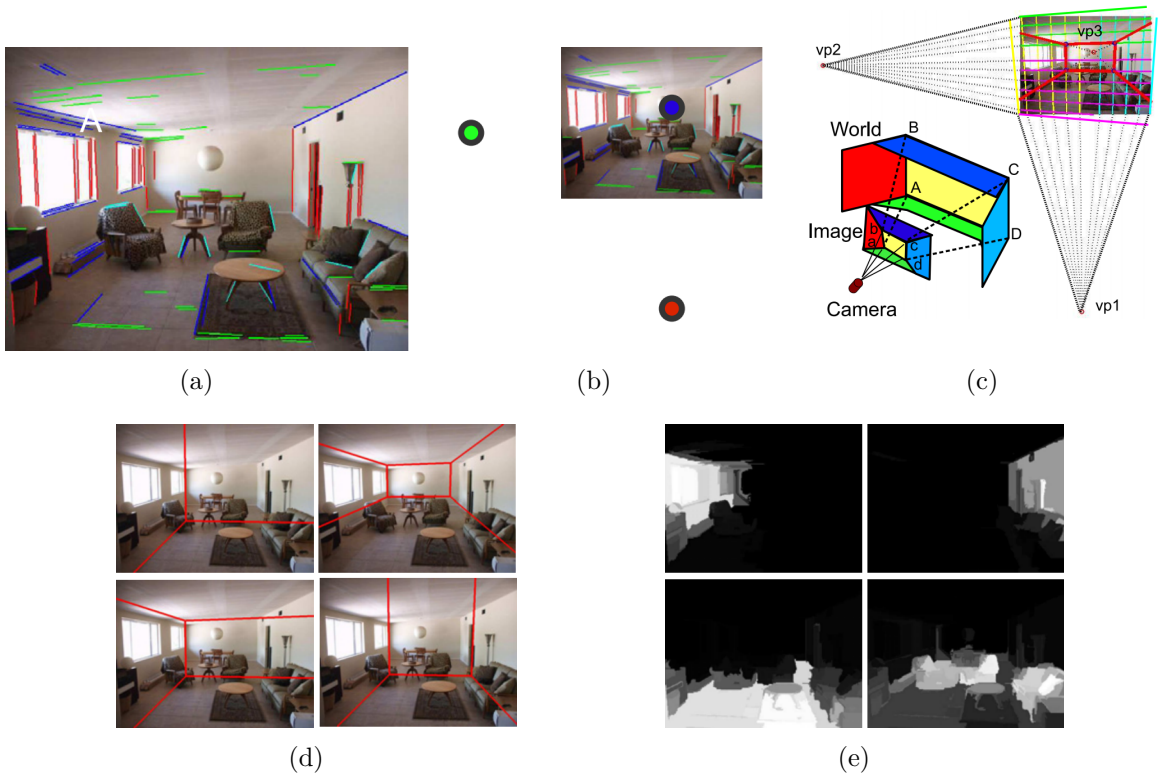


Figure 6.6 – Overview of the method of [Hedau et al., 2009]. (a) Lines are extracted and grouped in 3 main orthogonal directions: near vertical lines in red, near horizontal in green, depth oriented lines in blue, others in cyan, (b) Position of the 3 corresponding vanishing points, (c) Room layouts are generated by sampling rays originated from the vanishing points corresponding to vertical and horizontal lines, (d) Several sampled rooms, (e) Map of probabilities for labels "left wall", "right wall", "floor" and "object". This serves as features to rank the generated layouts. Figure from [Hedau et al., 2009].

feet positions. These votes are then summed up over the whole set of detections, which creates a "heat map" representing the empirical likelyhood for the feet position. We use this heat map to replace the ranking method from [Hedau et al., 2009] by the following scoring function f:

$$f(x, h, y) = \psi(x, y) + \alpha_\phi \phi(h, y) + \alpha_\rho \rho(y), \tag{6.1}$$

where x are image features, h is the estimated heat map, y is a room hypothesis and

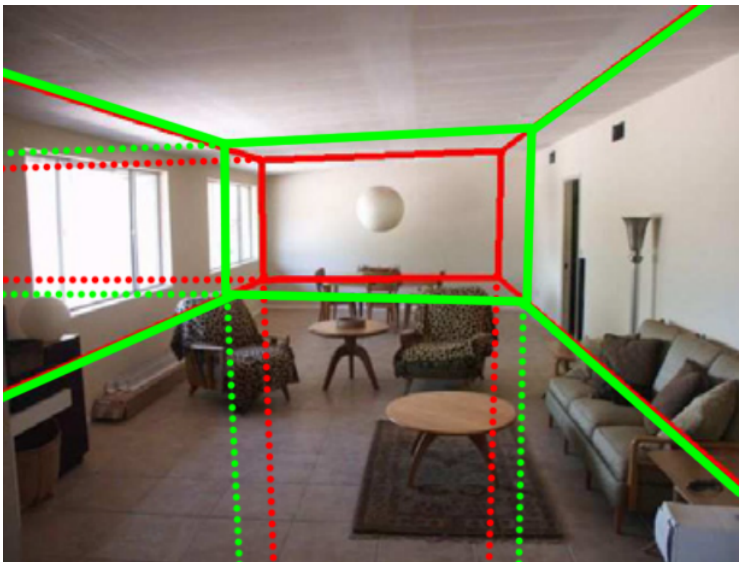


Figure 6.7 – Room layout hypotheses are generated by sampling rays from two out of the three vanishing points. Two successive rays are separated by a constant angle for each vanishing point. For deep rooms, when rays get close to the center of the back wall, a small angle difference between consecutive rays, e.g. red and green rays, can generate large differences in the room layout sizes. Here the depth between the red and green layouts probably differs by roughly one meter.

$\psi(x, y)$ is the original scoring method from [Hedau et al., 2009]. We add a term $\phi(h, y)$ which penalizes small rooms whose floors do not encompass the feet position heat map and a term $\rho(y)$ which penalizes large rooms. The parameters α_ϕ and α_ρ are cross-validated. Please refer to [Fouhey et al., 2012] for additional details.

6.3.3 Object localization (S3)

To assign each image pixel to an object label, we use the method described in Section 5.5.2 which results in a score representing how likely each pixel belongs to a specific label. In the following we refer to such maps of scores as "heat maps". The clutter heat map is the sum of scores for all the non-background classes (i.e. all classes except floor, wall or ceiling). We can either compute those heat maps by using appearance only (method (A+L) of the previous chapter, see Section 5.7) or by using both appearance and people (method (A+P)

of the previous chapter).

6.3.4 3D space occupancy (S4)

In the following, we aim at estimating the space occupied by the different object classes. We denote by $p = [p_x, p_y, p_z]^T$ the position of a 3D point in a coordinate system of the room, where p_z is the point's height above the floor. Given camera parameters K, R (see step S1) and c_i (see step S2), the projection $\pi_i(p)$ of any 3D point p for layout i can be computed by $\pi_i(p) = f(KR(p - c_i))$ where f transforms homogeneous coordinates into 2D pixel coordinates: $f(u) = [u_x/u_z; u_y/u_z]$. We discretize the 3D room space into voxels with 10 centimeters side. Our goal is to compute a score $s_l(p)$ that indicates how likely the voxel at 3D position p belongs to the l^{th} object label. The object label might either be one of the 7 object labels of the dataset (see Section 6.2) or background. We also compute a score indicating how likely the voxel belongs to "clutter" by summing the scores $s_l(p)$ for any non-background label l.

For the 3D clutter occupancy estimation, we follow [Hedau et al., 2009] and make the assumption that an occupied voxel must be supported by the floor. Its occupancy score is thus the sum of occupancy scores for the floor and for its projections in the image plane. See Figure 6.8(a) for an illustration. More formally, we assume we have a heat map $m_l([x; y])$ which stores the score representing how likely the pixel (x, y) belongs to a label l (see Section 6.3.3). Then we define $t_l(p) = m_l(\pi_i(p))$ which assigns a temporary score to each voxel given its projection in the image plane. This assigns the same score to all the voxels which project on the same pixel and a high score tends to leak away from the camera, see Figure 6.8(f). Assuming that an occupied voxel must be supported by the floor, the final score of each voxel is the sum of both a voxel and its ground supporting voxel temporary score: $s_l(p) = t_l(p) + t_l([p_x, p_y, 0]^T)$.

Knowing the height of objects, we can further refine the hypothesized 3D volume of an

object, see Figure 6.8(b) for an illustration. We compute the average object height h_l for each class l from the training data and define the score associated to a voxel to be $-\infty$ above this height or the sum of the heat map scores projected on the floor and at height h_l:

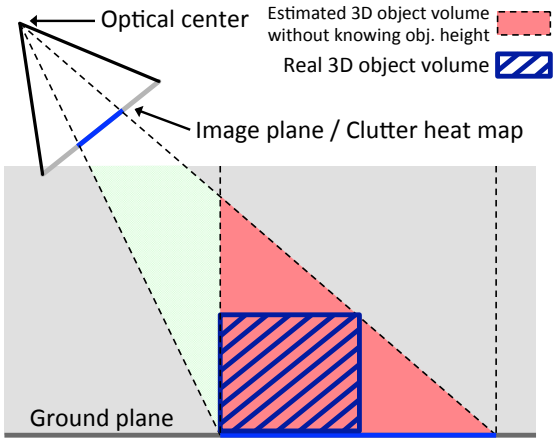
$$s_l(p) = \begin{cases} t_l\left([p_x, p_y, h_l]^T\right) + t_l\left([p_x, p_y, 0]^T\right) & \text{if } 0 \leq p_z \leq h_l \\ -\infty & \text{otherwise.} \end{cases} \tag{6.2}$$

One can also improve this step using the observation of people. By letting person detections vote around the feet position we create a heat map of person locations [Fouhey et al., 2012]. We normalize this heat map and compute its complement to 1, thus obtaining a likelihood of being "not walkable" that we project on the voxels of the floor, defining a score $\bar{w}(p_x, p_y)$. To represent the fact that no object can be present on a walkable area, we add the "not walkable" score to the previously defined temporary heat map t_l: $t_l(p) \leftarrow t_l(p) + \bar{w}(p_x, p_y)$.

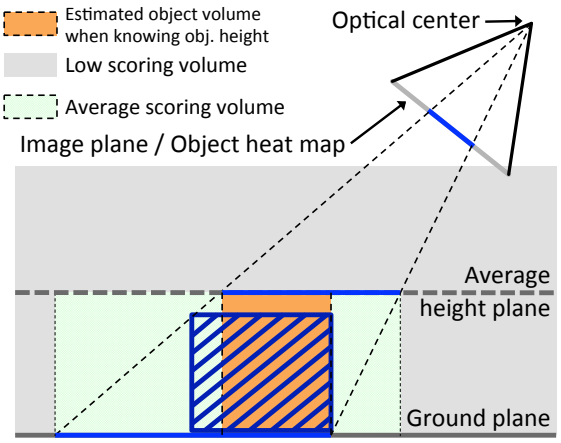
6.4 Performance measures

We first detail the different performance measures before presenting the experiments.

2D Acc: Most of the papers which address the problem of room layout estimation in Section 2.4 use the pixel-wise layout accuracy proposed by [Hedau et al., 2009]. In this measure, pixels are labeled as being either left wall, middle wall, right wall, floor or ceiling. Layouts are compared to ground truth by reporting the percentage of correct pixel labels. This measure is ambiguous in the case where only two walls are visible as they could be labeled either as "left+middle walls" or "middle+right walls". It is also not suited in cases where the four walls of the room are present in the image, for example when cameras are located in a corner of the ceiling. Finally large differences in 3D may be attenuated by

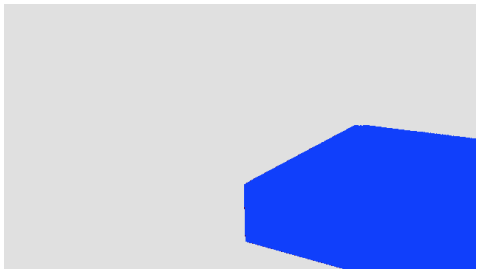


(a) Space occupancy scoring for clutter

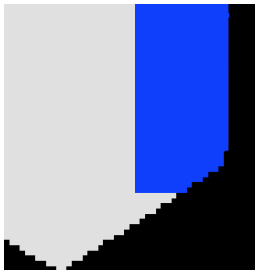


(b) Space occupancy scoring for object classes

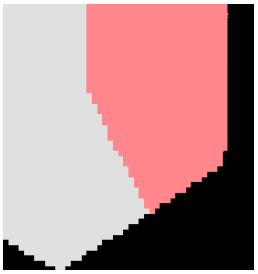
(c) Input image



(d) Ground truth for "Bed" (in blue)



(e) Top view of ground truth for "Bed" (in blue)



(f) Top view: floor-level space occupancy score computed with method (a)

(g) Top view: floor-level space occupancy score computed with method (b)

Figure 6.8 – Space estimation methods. (a) For clutter: we project the clutter heat map on the floor. The occupancy score of a voxel is the sum of the scores of its projections in the image plane and on the floor. (b) For an object class: we compute the average object height on the training set. The occupancy score of any voxel above this height is set to $-\infty$. The score of any other voxel is the sum of it projections on the floor and on the plane at the average height. (c-g) An example of estimated voxel scores. (c) Input image, (d) Ground truth heat map for the object class "Bed" (in blue), (e) Top-view of the ground truth map on the floor, (f) Top-view of the space occupancy score computed with the "clutter" method (a) at the floor level, (g) Top-view of the space occupancy score computed with the "object" method (b) at the floor level. Knowing the average height helps to better estimate the occupied space.

the projection in the image plane and the pixel-wise layout accuracy might not reflect the quality of the estimated layout [Hedau et al., 2012]. We therefore propose to use other measures of performance.

3D I/U: We follow [Chao et al., 2013] to compute how well the room volume has been estimated. We measure the intersection over union between the estimated and ground truth observable 3D spaces. We define the observable space as being the intersection between the room volume and the cone which goes from the camera center through the image. We place the camera center at the origin. As this measure is invariant by scaling, it is not impacted by the scale ambiguity inherent to monocular 3D calibration.

3D AP: Following [Hedau et al., 2012], we also estimate the occupied space average precision (AP). We discretize the room into 3D voxels and assign an occupancy score to each of them by following the method described in Section 6.3. We approximate objects by their 3D bounding boxes and convert their coordinates in the camera coordinate system. For a given object label, a voxel is labeled as positive if its center falls withing an object bounding box and as negative otherwise. This allows us to compute two types of average precision scores: (a) AP on the clutter by considering the space occupied by all object labels (**3D Clut. AP**) and (b) AP on the objects by considering only the space occupied by objects of a specific semantic class (**3D Obj. mAP**). To our knowledge, our work is the first to evaluate the estimation of semantic scape occupancy in 3D.

Pitch, Yaw, Roll: We also report the pitch, yaw and roll errors. These errors represent the difference of angle between the estimated and ground truth X, Y and Z axis of the camera, respectively.

6.5 Experiments

We first evaluate the performance of the method proposed for estimating the space occupancy (S4 - Space) when all other parameters are fixed to their ground truth (GT) values, i.e. camera calibration (S1 - Calib), layout selection (S2 - Layout) and object heat maps (S3 - Objects). The results are reported in the top row of Table 6.1 with (People) and without using people (Baseline). We can first notice the moderate performance as the baseline and the people augmented methods only reach an **3D AP** around 38% for both clutter and object evaluations. This can be explained by the fact that our modeling of clutter tends to produce high scoring cones of voxels, see Figure 6.8(f). On the contrary, our modeling of objects avoids this issue but relies on the average height of objects which can be quite variable and could be inaccurate. We also see that using people does not improve the average precision in this setup. This is understandable as we only use people to remove false positive clutter/object scores on walkable areas: as we already have a perfect object label, people can only act as a source of noise. The **3D AP** performance per object label for this setup is detailed in the first column of Table 6.2.

In a more realistic situation where we estimate the 2D object heat maps automatically, the use of people improves average precision for both clutter and objects (second row of Table 6.1). Some good qualitative results for row 2 are depicted in the third column of Figure 6.9. Although the shape of objects is not accurately captured, the space occupied by different semantic classes of objects can be relatively well estimated. The details of the per-class **3D AP** are shown in the second column of Table 6.2.

In the following we stop using the ground truth 3D room layout and try to estimate it. The situation where no ground truth is used at all is described by the last row of Table 6.1. We can see a huge performance drop in average precision for space occupancy estimation that we try to explain in the next paragraph. Some examples of failures are shown in

Figure 6.10. Examples of success are depicted in the fourth column of Figure 6.9. This is the only setup where the camera parameters are estimated. We report rotation errors of 7.7°, 5.5° and 10.1° for pitch, yaw and roll, respectively. This is significantly higher than the results obtained by the same algorithm on the dataset of [Chao et al., 2013]. The **3D AP** per object class is shown in the last column of Table 6.2.

In order to identify the source of confusion we evaluated the performance with ground truth calibration but the gain in performance (row 4 of Table 6.1) is negligible compared to the previous setup. In order to further investigate this, we propose a setup where we select the best layout according to the **3D I/U** score among all the candidate layouts proposed by the algorithm from [Hedau et al., 2009]. The results are shown in the third row of Table 6.1. The **3D I/U** score is an upper bound of what is achievable with this algorithm and this average result reveals this stage of the pipeline is not able to generate the correct layouts despite the angle step reduction mentioned in Subsection 6.3.2. This results in a large noise in the estimation of the camera position c_i (see Subsection 6.3.2) which might explain why the AP is very low as our method to estimate the occupied space critically relies on the projection of the object heat maps on the floor. Especially, if the camera height is not precisely estimated the projected heat maps are shifted away from (height overestimated) or towards (height underestimated) the camera. To visualize this, please refer to Figure 6.8(a) and imagine what happens to the blue area on the floor when the camera moves up or down while keeping its orientation fixed. Please also note how the **3D I/U** score is multiplied by more than two whereas the 2D pixel layout accuracy **2D Acc** goes down between rows 4 and 3 in Table 6.1. This confirms that the **2D Acc** measure is not indicative of the 3D errors as pointed out by [Hedau et al., 2012].

	S1 Calib	S2 Layout	S3 Objects	S4 Space	2D Acc	3D I/U	3D Clut. AP	3D Obj. mAP
1	GT			People	100±0.0	100±0.0	**38.3±1.6**	37.6±6.3
				Baseline			37.5±1.4	**38.6±5.9**
2	GT		People		100±0.0	100±0.0	**36.6±2.7**	**22.2±4.6**
			Baseline				32.2±2.4	20.0±3.3
3	GT	Best estimated	People		66.1±1.7	66.1±2.8	**16.0±3.2**	**8.5±3.0**
			Baseline				14.7±2.6	8.0±2.9
4	GT	People			**74.2±1.7**	**30.7±3.7**	**5.2±1.2**	**1.9±1.4**
		Baseline			71.7±1.6	30.6±2.6	4.9±1.7	**1.9±1.1**
5	Hedau09	People			**72.9±2.3**	**32.9±3.2**	5.0±1.5	1.8±1.6
		Baseline			71.1±0.9	31.2±3.4	**5.1±1.8**	**1.9±1.8**

Table 6.1 – All the numbers are percentages. Results of the room layout estimation 2D accuracy (2D Acc) and 3D intersection over union (3D I/U), 3D clutter average precision (3D Clut. AP) and 3D object mean average precision (3D Obj. mAP). Please refer to the text for the discussion.

	1		2		5	
S1 - Calibration	GT	GT	GT	GT	Hedau	Hedau
S2 - Layout	GT	GT	GT	GT	People	Baseline
S3 - Object	GT	GT	People	Baseline	People	Baseline
S4 - Space	People	Baseline	People	Baseline	People	Baseline
Bed	**53±3.8**	49±2.8	**48±7.7**	45±4.9	2.6±3.7	**2.9±4.5**
Sofa/Armchair	45±3.5	**48±5.4**	**33±4.6**	30±3.9	**2.8±1.9**	2.4±0.8
Coffee Table	**47±11.7**	46±7.7	**20±5.7**	13±2.6	**0.9±1.3**	0.3±0.2
Chair	26±5.5	**31±5.4**	7.0±2.3	**8.4±2.3**	**1.1±0.7**	0.5±0.4
Table	44±3.9	**45±4.0**	**19±2.8**	17±2.9	**1.8±1.2**	1.7±0.9
Cupboard	**32±3.2**	31±2.9	**18±1.8**	16±1.9	2.2±1.2	**3.0±2.1**
Christmas tree	38±16	**44±16**	**29±11**	26±6.8	1.7±2.1	**3.9±4.9**
Other Object	**15±2.8**	**15±2.6**	**4.5±1.4**	4.4±1.4	**1.0±0.8**	0.8±0.5
Average	**37.6±6.3**	**38.6±5.9**	**22.2±4.6**	20.0±3.3	**1.8±1.6**	1.9±1.8

Table 6.2 – Detailed **3D AP** per object class. All the numbers are percentages. Average precision for each object class. The three different columns correspond to rows 1, 2 and 5 of Table 6.1, respectively. Column 2 corresponds to the setup where we expect people to improve over the baseline and we see that it indeed performs slightly better.

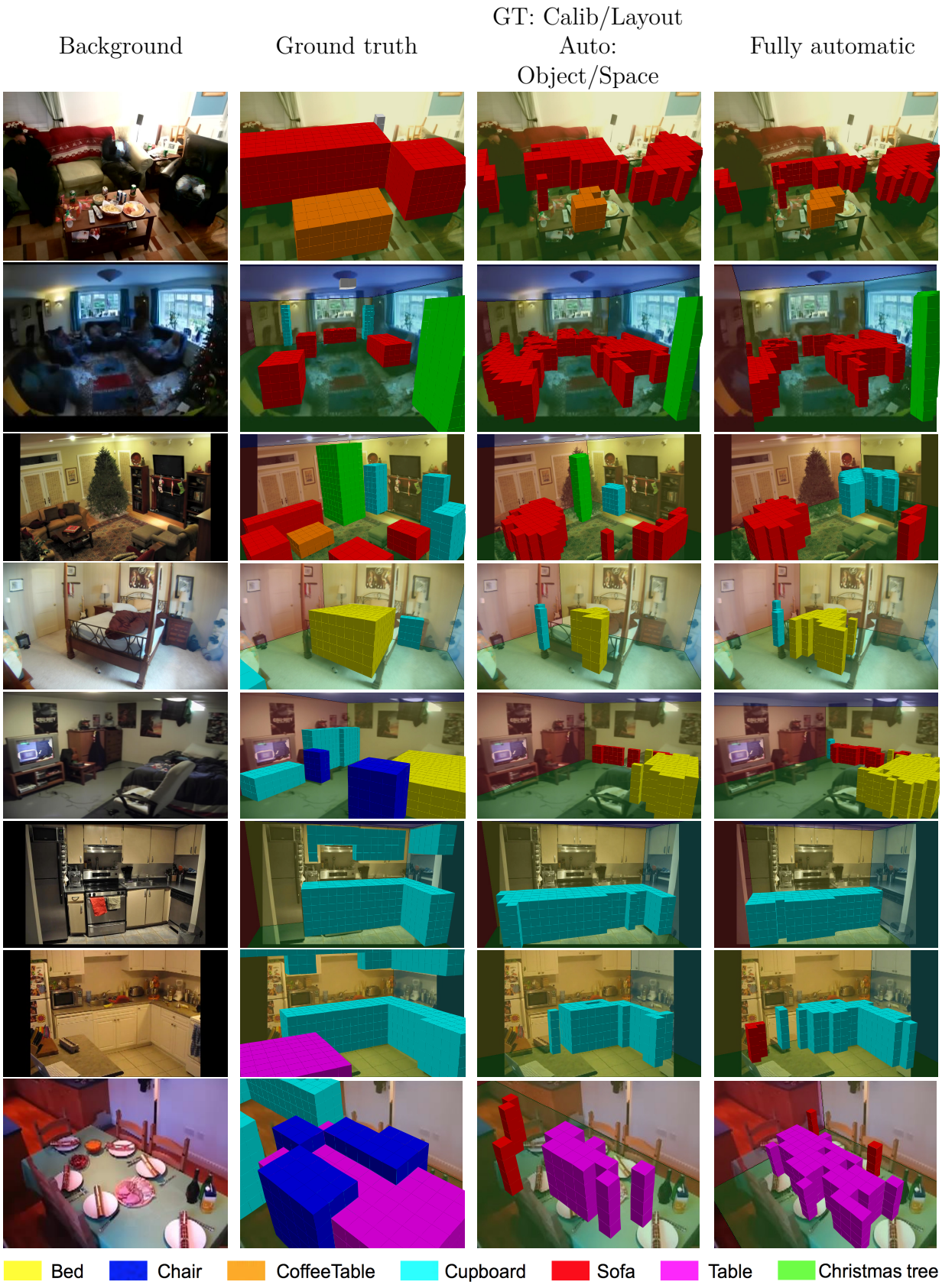


Figure 6.9 – Semantic 3D space occupancy estimation. Columns from left to right: room background, ground truth 3D semantic labeling, results corresponding to second row (People) of Table 6.1 and results corresponding to the last row (People) of Table 6.1.

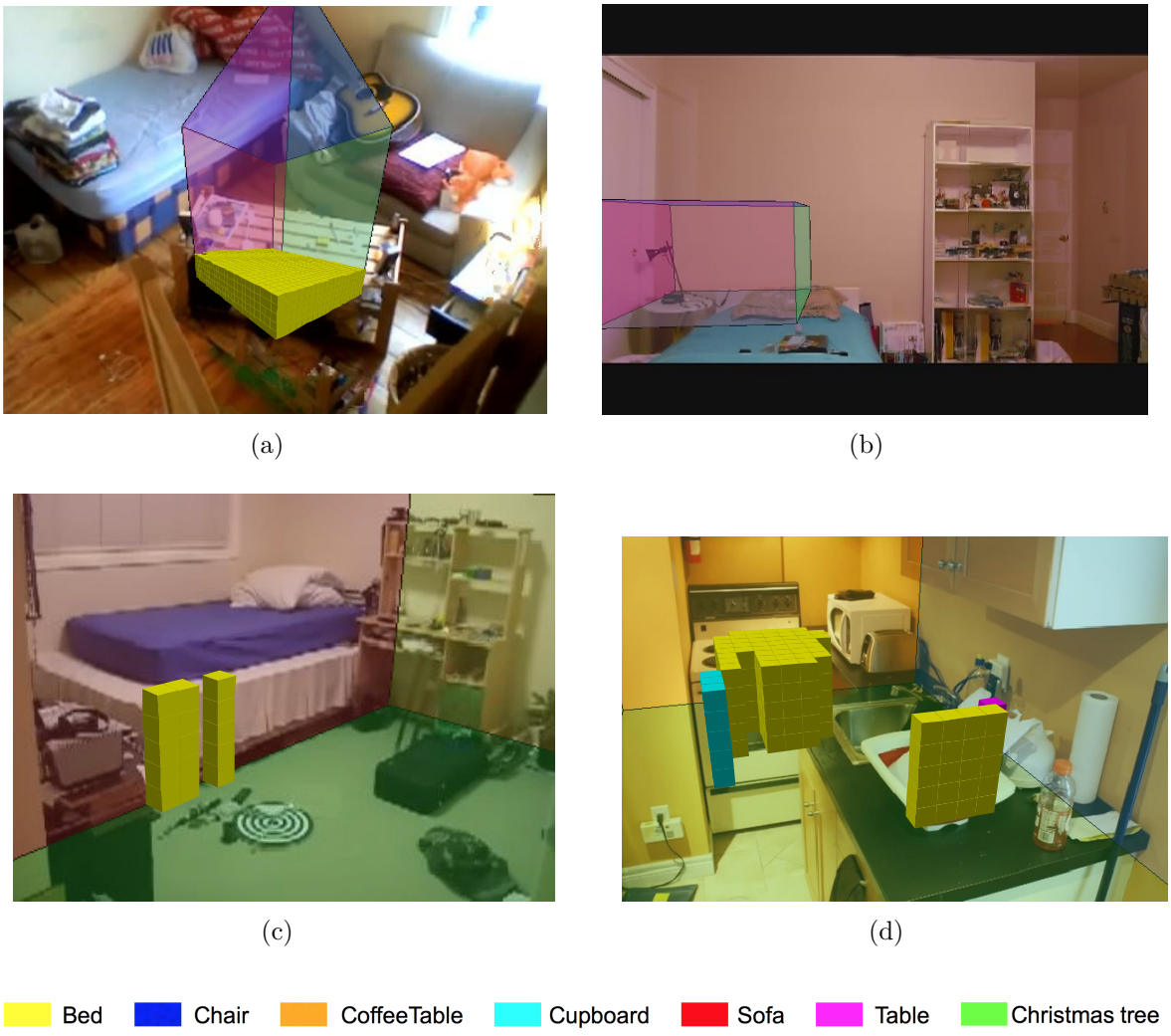


Figure 6.10 – Four different causes of failure: (a) vanishing point estimation fails, (b) layout estimation completely fails, (c) layout estimation fails and incorrectly removes most of the objects in the room, (d) object label estimation fails.

6.6 Discussion

We have presented a dataset with new 3D ground truth annotations of the room layout and semantic object volumes. This allowed us to evaluate the 3D semantic occupancy estimation of the room. We have shown that the use of person cues extracted from dynamic scenes with acting people allows us to obtain better performance in many of the experimental setups. However, overall, the standard pipeline performs poorly on the estimation of 3D occupied space when using no ground truth data, see last row of Table 6.1 and last column of Table 6.2. To understand this low performance, we evaluated the impact of using ground truth annotations at different stages of the layout estimation pipeline. We have shown that the most critical part is the selection of the 3D room layout. The candidate layouts appear to be too coarsely sampled, especially for deep rooms when the sampling of rays coming from the vanishing points is not precise enough to generate the right candidate, as discussed in Figure 6.7. In addition, the box model might not be suited for all indoor scenes. For example, some rooms in our dataset have the shape of a "L" and their layout is not well approximated by a box. One can thus question if a box is an appropriate model for realistic rooms. This argues in favor of a performance measure more representative than the 2D pixel accuracy, which is only suited for box layouts. Alternative measures include the free floor estimation used in [Satkin et al., 2012], which only evaluates top view, and the 3D semantic object volume overlap used in this chapter.

Chapter 7

Discussion

In this chapter, we summarize contributions of the thesis and discuss future work.

7.1 Contributions of the thesis

This thesis has focused on the interactions between people, objects and scenes. Motivated by the continuously growing amount of data depicting people and their environment as well as by psychological studies suggesting a tight coupling between the way we recognize people's actions and their context, we have demonstrated how modeling such coupling can lead to improvements in action classification and scene understanding.

7.1.1 Action classification

Recognizing actions is a very difficult task as people may accomplish the same activity in many different ways. Automatic estimation of human poses could help this task. Reliable human pose estimation, however, is a challenge by itself as the projection of the 3D plausible poses can lead to numerous and ambiguous 2D configurations. We were thus interested

in developing methods to capture other sources of information for action classification and we have investigated the modeling of scene and object context. In Chapter 3, we realized a study of the bag-of-features model applied to action classification. We manually collected a new dataset for action classification in still images. We have investigated different experimental setups with different kernels and vocabulary sizes and showed that the best performing combination compared favorably to the deformable part-based model (DPM) of [Felzenszwalb et al., 2009] on three different datasets. We also demonstrated the positive impact of scene context on the performance by using different kernels computed on the person and scene background separately. To further investigate the effect of context, we proposed in Chapter 4 a new image descriptor to capture interactions between body parts and objects. Instead of describing an image with a bag of HOG or SIFT features, our descriptor relies on the relative displacement of pairs of discriminative body parts or object detectors over scale and space. We showed how to generate a pool of candidate pairs and how to remove redundant or irrelevant features using a sparsity inducing regularizer for discriminative feature selection. For the task of action recognition in still images, we outperformed the strong bag-of-features baseline and, combined with the DPM, obtained performance close to the state of the art.

7.1.2 Scene understanding

Localizing objects and understanding the layout of cluttered indoor scenes are also difficult tasks. Heavy occlusions of objects and multiplicity of viewpoints are difficult to handle and state-of-the-art models still perform poorly on such data. This is why we have investigated the use of people as an additional source of information to help the understanding of indoor scenes. In Chapter 5, we relied on people to improve localization of objects in a room. We gathered a new dataset of indoor time-lapses captured by a static camera and used the method of [Yang and Ramanan, 2011] to detect people and estimate their poses.

We developed a pose-based approach to describe an image region by the distribution of human poses around it. Using our dataset, we demonstrated that using such pose cues combined with visual cues significantly improves over the SIFT-based bag-of-features or DPM baselines. We have developed a visualization and qualitative interpretation of the parameters learned by our model. We used this interpretation to propose a baseline method to generate plausible poses from a semantic segmentation of the room, showing the benefit of the link between context and human pose. In Chapter 6, we extended the annotation of our time-lapse dataset to 3D in order to investigate the possibility of using people to improve the 3D semantic scene understanding. We extended the standard four-step pipeline (i.e. camera calibration, room layout selection, 2D object localization and semantic 3D space occupancy) to take advantage of people in the three last steps. We evaluated the performance of the method for 3D semantic space occupation estimation and showed that people could lead to significant improvements when the room layout has been correctly estimated. However we also demonstrated that the layout estimation is currently the most critical step of the pipeline. Using our ground truth annotation, we showed that the best performing layouts generated by the method of [Hedau et al., 2009] are still not precise enough for the subsequent parts of the pipeline to produce good results.

7.2 Future work

In this section we discuss possible directions for future research.

7.2.1 Action classification

Using pose features: Pose estimation has seen a recent boost in performance with the advances in deep learning. The method proposed in [Tompson et al., 2014] for example achieves a high detection rate and a low pose estimation error. One can thus now consider

using poses as a more reliable feature for action classification. This raises interesting representation and learning questions as the L2-norm on 2D joint positions, as we have used in Chapter 5, is probably not the best way to compute similarity between two poses. Possible directions of research thus include finding an appropriate representation of 2D poses which would be invariant under small view-point changes and finding the metric which is the best suited to compare pose features.

Hierarchical models: The state of the art action classifier/detectors can address the problem of occlusions, see for example the model of [Desai and Ramanan, 2012] which explicitly deals with occluded parts. However such deformable part models are organized in a tree and an occluded part might prevent predicting the position of its children accurately. The pose estimation method of [Sapp and Taskar, 2013] might solve such issue. They propose a hierarchical approach with large scale parts covering more than one body part, e.g. covering a full arm. The advantage for action classification could be two fold: (i) higher level parts would better capture context, and (ii) be less sensitive to occlusions.

Modeling the mutual context: The development of unsupervised learning of mid-level patches [Singh et al., 2012, Sun and Ponce, 2013] could allow to automatically learn a set of discriminative patches for poselet-like body parts, object parts and scene parts. This would provide more reliable features to extend the method developed in Chapter 4 for learning interactions between people, objects and scenes. One may also improve this method by using a non-linear classifier such as a random forest or group mid-level detectors in inference trees by learning a set of Chow-Liu trees [Chow and Liu, 1968].

7.2.2 Scene understanding

3D poses: One could extend the work of Chapter 5 to 3D by back-projecting 2D poses to the 3D scene, for example assuming each pose has to touch the ground. One could then divide space around joints into volume regions in the same manner as proposed in Section 5.3. This would allow us to define pose features on voxels instead of image regions. One could also define additional features on voxels based on: (a) the back-projection of 2D object heat maps obtained in Chapter 5, (b) the voxel height (c) the voxel distance to its closest wall. The label of each voxel of a room could then be predicted at once using a structured SVM. Another possible direction of work would be to extend our work from [Fouhey et al., 2012]. By projecting the 2D detected poses to 3D, one could start reasoning about the ratio between people's height and room's height which would give a prior on the room size.

Use context: The proposed method in Chapter 6 does not reason about objects in terms of instances but just in terms of 3D volumes. One cannot properly use context with this representation as it cannot model the orientation of objects. This could be useful, as a coffee table is often placed in front of a sofa for example. It does not model the position of a voxel within an object instance neither, which could be of interest as, for example, a bottom voxel of a TV must be supported by a top voxel of a cupboard or a table. A possible direction for future work could be to extend the model of Chapter 6 to refine a set of object 3D bounding box hypotheses.

Changing the room model: We have shown in the last chapter that approaches selecting the room layout before reasoning about objects in the scene are not likely to work well in cluttered scenes. More generally, the gap between estimating a room layout and detecting 3D object instances in a room is too big and we may need to build better tools to

reason locally about objects in rooms before making any decision about the global interpretation. This has been pointed out by [Fouhey et al., 2014] who propose a state-of-the art method to assign a surface normal to each pixel of an image in a constrained manner. This seems to be a good prior to build on for 3D object boxes, supporting planes and occlusion boundaries estimation in realistic images of cluttered indoor scenes.

Bibliography

[Agarwal and Triggs, 2004] Agarwal, A. and Triggs, B. (2004). 3d human pose from silhouettes by relevance vector regression. In CVPR.

[Alexe et al., 2010] Alexe, B., Deselaers, T., and Ferrari, V. (2010). What is an object? In CVPR.

[Andriluka et al., 2009] Andriluka, M., Roth, S., and Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In CVPR.

[Andriluka et al., 2010] Andriluka, M., Roth, S., and Schiele, B. (2010). Monocular 3D pose estimation and tracking by detection. In CVPR.

[Bach et al., 2005] Bach, P., Knoblich, G., Gunter, T. C., Friederici, A. D., and Prinz, W. (2005). Action comprehension: deriving spatial and functional relations. Journal of Experimental Psychology: Human Perception and Performance, 31:465.

[Bao et al., 2011] Bao, S. Y., Sun, M., and Savarese, S. (2011). Toward coherent object detection and scene layout understanding. Image and Vision Computing, 29:569–579.

[Barinova et al., 2010] Barinova, O., Lempitsky, V., Tretyak, E., and Kohli, P. (2010). Geometric image parsing in man-made environments. In ECCV.

[Bobick and Davis, 2001] Bobick, A. and Davis, J. (2001). The recognition of human movement using temporal templates. PAMI.

[Bosch et al., 2007] Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In Proc. CIVR.

[Bourdev et al., 2010] Bourdev, L., Maji, S., Brox, T., and Malik, J. (2010). Detecting people using mutually consistent poselet activations. In ECCV.

[Bourdev and Malik, 2009] Bourdev, L. and Malik, J. (2009). Poselets: Body part detectors trained using 3D human pose annotations. In ICCV.

[Brox et al., 2011] Brox, T., Bourdev, L., Maji, S., and Malik, J. (2011). Object segmentation by alignment of poselet activations to image contours. In CVPR.

[Bub and Masson, 2006] Bub, D. and Masson, M. (2006). Gestural knowledge evoked by objects as part of conceptual representations. Aphasiology, 20:1112–1124.

[Chao and Martin, 2000] Chao, L. L. and Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. Neuroimage, 12:478–484.

[Chao et al., 2013] Chao, Y.-W., Choi, W., Pantofaru, C., and Savarese, S. (2013). Layout estimation of highly cluttered indoor scenes using geometric and semantic cues. In ICIAP.

[Choi et al., 2013] Choi, W., Chao, Y.-W., Pantofaru, C., and Savarese, S. (2013). Understanding indoor scenes using 3d geometric phrases. In CVPR.

[Chow and Liu, 1968] Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on, 14:462–467.

[Csurka et al., 2004] Csurka, G., Bray, C., Dance, C., and Fan, L. (2004). Visual categorization with bags of keypoints. In WS-SLCV, ECCV.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In CVPR.

[Dantone et al., 2013] Dantone, M., Gall, J., Leistner, C., and Van Gool, L. (2013). Human pose estimation using body parts dependent joint regressors. In CVPR.

[Dean et al., 2013] Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S., and Yagnik, J. (2013). Fast, accurate detection of 100,000 object classes on a single machine. In CVPR.

[Del Pero et al., 2012] Del Pero, L., Bowdish, J., Fried, D., Kermgard, B., Hartley, E. L., and Barnard, K. (2012). Bayesian geometric modeling of indoor scenes. In CVPR.

[Del Pero et al., 2011] Del Pero, L., Guan, J., Brau, E., Schlecht, J., and Barnard, K. (2011). Sampling bedrooms. In CVPR.

[Delaitre et al., 2012] Delaitre, V., Fouhey, D., Laptev, I., Sivic, J., Efros, A. A., and Gupta, A. (2012). Scene semantics from long-term observation of people. In ECCV.

[Delaitre et al., 2010a] Delaitre, V., Laptev, I., and Sivic, J. (2010a). Recognizing human actions in still images: a study of bag-of-features and part-based representations. In Proc. BMVC. updated version, available at http://www.di.ens.fr/willow/research/stillactions/.

[Delaitre et al., 2010b] Delaitre, V., Laptev, I., and Sivic, J. (2010b). Willow actions database. http://www.di.ens.fr/willow/research/stillactions/.

[Delaitre et al., 2011] Delaitre, V., Sivic, J., and Laptev, I. (2011). Learning person-object interactions for action recognition in still images. In NIPS.

[Desai and Ramanan, 2012] Desai, C. and Ramanan, D. (2012). Detecting actions, poses, and objects with relational phraselets. In ECCV.

[Desai et al., 2010] Desai, C., Ramanan, D., and Fowlkes, C. (2010). Discriminative models for static human-object interactions. In CVPR.

[Deutscher et al., 2000] Deutscher, J., Blake, A., and Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In CVPR.

[Deutscher et al., 1999] Deutscher, J., North, B., Bascle, B., and Blake, A. (1999). Tracking through singularities and discontinuities by random sampling. In ICCV.

[Doersch et al., 2013] Doersch, C., Gupta, A., and Efros, A. A. (2013). Mid-level visual element discovery as discriminative mode seeking. In NIPS.

[Doersch et al., 2012] Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2012). What makes paris look like paris? In SIGGRAPH.

[Dollár et al., 2005] Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In ICCV Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance.

[Elgammal and Lee, 2004] Elgammal, A. and Lee, C.-S. (2004). Inferring 3d body pose from silhouettes using activity manifold learning. In CVPR.

[Everingham et al., 2007] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010.

[Everingham et al., 2012] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012.

[Fathi et al., 2011] Fathi, A., Ren, X., and Rehg, J. (2011). Learning to recognize objects in egocentric activities. In CVPR.

[Fei-Fei and Li, 2010] Fei-Fei, L. and Li, L.-J. (2010). What, where and who? telling the story of an image by activity classification, scene recognition and object categorization. In Computer Vision, pages 157–171. Springer.

[Fei-Fei and Perona, 2005] Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In CVPR.

[Felzenszwalb, 2001] Felzenszwalb, P. (2001). Learning models for object recognition. In CVPR.

[Felzenszwalb et al., 2009] Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part based models. PAMI.

[Felzenszwalb and Huttenlocher, 2004] Felzenszwalb, P. and Huttenlocher, D. (2004). Distance transforms of sampled functions. Technical report, Cornell University CIS, Tech. Rep. 2004-1963.

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

[Felzenszwalb and Huttenlocher, 2000] Felzenszwalb, P. and Huttenlocher, D. P. (2000). Efficient matching of pictorial structures. In CVPR.

[Felzenszwalb et al., 2008a] Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008a). Discriminatively trained deformable part models. http://www.cs.berkeley.edu/r̃bg/latent/.

[Felzenszwalb et al., 2008b] Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008b). A discriminatively trained, multiscale, deformable part model. In CVPR.

[Ferrari et al., 2008a] Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008a). Progressive search space reduction for human pose estimation. In CVPR.

[Ferrari et al., 2008b] Ferrari, V., Marin-Jimenez, M., and Zisserman, A. (2008b). Progressive search space reduction for human pose estimation. CVPR.

[Fischler and Elschlager, 1973] Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. IEEE Transactions on Computer, 22:67–92.

[Fouhey et al., 2012] Fouhey, D. F., Delaitre, V., Gupta, A., Efros, A. A., Laptev, I., and Sivic, J. (2012). People watching: Human actions as a cue for single-view geometry. In ECCV.

[Fouhey et al., 2013] Fouhey, D. F., Gupta, A., and Hebert, M. (2013). Data-driven 3d primitives for single image understanding. In ICCV.

[Fouhey et al., 2014] Fouhey, D. F., Gupta, A., and Hebert, M. (2014). Unfolding an indoor origami world. In ECCV.

[Freund and Schapire, 1997] Freund, Y. and Schapire, R. (1997). A decision theoretic generalisation of online learning. Computer and System Sciences, 55:119–139.

[Gall et al., 2011] Gall, J., Fossati, A., and van Gool, L. (2011). Functional categorization of objects using real-time markerless motion capture. In CVPR.

[Gallese et al., 1996] Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119:593–609.

[Gallese and Goldman, 1998] Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in cognitive sciences, 2:493–501.

[Gavrila, 2000] Gavrila, D. M. (2000). Pedestrian detection from a moving vehicle. In ECCV.

[Geiger et al., 2011] Geiger, A., Wojek, C., and Urtasun, R. (2011). Joint 3d estimation of objects and scene layout. In NIPS.

[Gibson, 1979] Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

[Girshick et al., 2011] Girshick, R., Felzenszwalb, P., and McAllester, D. (2011). Object detection with grammar models. In NIPS.

[Gkioxari et al., 2013] Gkioxari, G., Arbeláez, P., Bourdev, L., and Malik, J. (2013). Articulated pose estimation using discriminative armlet classifiers. In CVPR.

[Gordon et al., 1993] Gordon, N. J., Salmond, D. J., and Smith, A. F. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. In IEE Proceedings F (Radar and Signal Processing).

[Gorelick et al., 2007] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. PAMI.

[Grabner et al., 2011] Grabner, H., Gall, J., and van Gool, L. (2011). What makes a chair a chair? In CVPR.

[Grochow et al., 2004] Grochow, K., Martin, S. L., Hertzmann, A., and Popović, Z. (2004). Style-based inverse kinematics. In SIGGRAPH.

[Gupta et al., 2008a] Gupta, A., Chen, T., Chen, F., Kimber, D., and Davis, L. (2008a). Context and observation driven latent variable model for human pose estimation. In CVPR.

[Gupta et al., 2010] Gupta, A., Efros, A. A., and Hebert, M. (2010). Blocks world revisited: Image understanding using qualitative geometry and mechanics. In ECCV.

[Gupta et al., 2009] Gupta, A., Kembhavi, A., and Davis, L. S. (2009). Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI.

[Gupta et al., 2008b] Gupta, A., Mittal, A., and Davis, L. S. (2008b). Constraint integration for efficient multiview pose estimation with self-occlusions. PAMI, 30:493–506.

[Gupta et al., 2011] Gupta, A., Satkin, S., Efros, A. A., and Hebert, M. (2011). From 3D scene geometry to human workspace. In CVPR.

[Hall et al., 2000] Hall, D., de Verdière, V. C., and Crowley, J. L. (2000). Object recognition using coloured receptive fields. In ECCV.

[Hara and Chellappa, 2013] Hara, K. and Chellappa, R. (2013). Computationally efficient regression on a dependency graph for human pose estimation. In CVPR.

[Harzallah et al., 2009] Harzallah, H., Jurie, F., and Schmid, C. (2009). Combining efficient object localization and image classification. In ICCV.

[Hastie et al., 2003] Hastie, T., Tibshirani, R., and Friedman, J. H. (2003). The Elements of Statistical Learning. Springer.

[Hedau et al., 2009] Hedau, V., Hoiem, D., and Forsyth, D. (2009). Recovering the spatial layout of cluttered rooms. In ICCV.

[Hedau et al., 2010] Hedau, V., Hoiem, D., and Forsyth, D. (2010). Thinking inside the box: Using appearance models and context based on room geometry. In ECCV.

[Hedau et al., 2012] Hedau, V., Hoiem, D., and Forsyth, D. (2012). Recovering free space of indoor scenes from a single image. In CVPR.

[Helbig et al., 2006] Helbig, H. B., Graf, M., and Kiefer, M. (2006). The role of action representations in visual object recognition. Experimental Brain Research, 174:221–228.

[Hoiem et al., 2005] Hoiem, D., Efros, A. A., and Hebert, M. (2005). Geometric context from a single image. In ICCV.

[Hou et al., 2007] Hou, S., Galata, A., Caillette, F., Thacker, N., and Bromiley, P. (2007). Real-time body tracking using a gaussian process latent variable model. In ICCV.

[Ikizler et al., 2008] Ikizler, N., Cinbis, R. G., Pehlivan, S., and Duygulu, P. (2008). Recognizing actions from still images. In Proc. ICPR.

[Ikizler et al., 2009] Ikizler, N., Cinbis, R. G., and Sclaroff, S. (2009). Learning actions from the Web. In ICCV.

[Ioffe and Forsyth, 2001] Ioffe, S. and Forsyth, D. A. (2001). Probabilistic methods for finding people. IJCV.

[Isard and Blake, 1998] Isard, M. and Blake, A. (1998). Condensation—conditional density propagation for visual tracking. IJCV, 29:5–28.

[Jabri et al., 2000] Jabri, S., Duric, Z., Wechsler, H., and Rosenfeld, A. (2000). Detection and location of people in video images using adaptive fusion of color and edge information. In Proc. ICPR.

[Jhuang et al., 2007] Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In ICCV.

[Johnson, 2010] Johnson, S. (2010). Leeds sports pose dataset. http://www.comp.leeds.ac.uk/mat4saj/lsp.html.

[Johnson and Everingham, 2010] Johnson, S. and Everingham, M. (2010). Clustered pose and nonlinear appearance models for human pose estimation. In Proc. BMVC.

[Johnson and Everingham, 2011] Johnson, S. and Everingham, M. (2011). Learning effective human pose estimation from inaccurate annotation. In CVPR.

[Johnson-Frey et al., 2003] Johnson-Frey, S. H., Maloof, F. R., Newman-Norlund, R., Farrer, C., Inati, S., and Grafton, S. T. (2003). Actions or hand-object interactions? human inferior frontal cortex and action observation. Neuron, 39:1053–1058.

[Kanaujia et al., 2007] Kanaujia, A., Sminchisescu, C., and Metaxas, D. (2007). Spectral latent variable models for perceptual inference. In ICCV.

[Khan et al., 2014] Khan, S. H., Bennamoun, M., Sohel, F., and Togneri, R. (2014). Geometry driven semantic labeling of indoor scenes. In ECCV.

[Kitani et al., 2012] Kitani, K. M., Ziebart, B. D., Bagnell, J. A., and Hebert, M. (2012). Activity forecasting. In ECCV.

[Kjellstrom et al., 2008] Kjellstrom, H., Romero, J., Martinez, D., and Kragic, D. (2008). Simultaneous visual recognition of manipulation actions and manipulated objects. In ECCV.

[Kohli and Torr, 2009] Kohli, P. and Torr, P. (2009). Robust higher order potentials for enforcing label consistency. IJCV, 82:302–324.

[Kourtzi, 2004] Kourtzi, Z. (2004). But still, it moves. Trends in cognitive sciences, 8:47–49.

[Kourtzi and Kanwisher, 2000] Kourtzi, Z. and Kanwisher, N. (2000). Activation in human mt/mst by static images with implied motion. Journal of cognitive neuroscience, 12:48–55.

[Krahnstoever and Mendonca, 2005] Krahnstoever, N. and Mendonca, P. R. (2005). Bayesian autocalibration for surveillance. In ICCV.

[Kuehne et al., 2011] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). HMDB: a large video database for human motion recognition. In ICCV.

[Laptev, 2013] Laptev, I. (2013). Modeling and visual recognition of human actions and interactions. Habilitation à diriger des recherches en mathématiques et en informatique, École normale supérieure, Paris, France.

[Laptev and Lindeberg, 2003] Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In ICCV.

[Laptev et al., 2008] Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In CVPR.

[Laptev and Pérez, 2007] Laptev, I. and Pérez, P. (2007). Retrieving actions in movies. In ICCV.

[Lawrence, 2003] Lawrence, D. N. (2003). Gaussian process latent variable models for visualisation of high dimensional data. In NIPS.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In CVPR.

[Lee et al., 2010] Lee, D., Gupta, A., Hebert, M., and Kanade, T. (2010). Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In NIPS.

[Lee et al., 2009] Lee, D., Hebert, M., and Kanade, T. (2009). Geometric reasoning for single image structure recovery. In ICCV.

[Li et al., 2010] Li, L., Su, H., Xing, E., and Fei-Fei, L. (2010). Object bank: A high-level image representation for scene classification and semantic feature sparsification. In NIPS.

[Li and Fei-Fei, 2007] Li, L. J. and Fei-Fei, L. (2007). What, where and who? Classifying events by scene and object recognition. In ICCV.

[Lowe, 1999] Lowe, D. (1999). Object recognition from local scale-invariant features. In ICCV.

[Lowe, 2004] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. IJCV, 60:91–110.

[Maji et al., 2011] Maji, S., Bourdev, L., and Malik, J. (2011). Action recognition from a distributed representation of pose and appearance. In CVPR.

[Marszalek et al., 2009] Marszalek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In CVPR.

[Mikolajczyk et al., 2004] Mikolajczyk, K., Schmid, C., and Zisserman, A. (2004). Human detection based on a probabilistic assembly of robust part detectors. In ECCV.

[Mohan et al., 2001] Mohan, A., Papageorgiou, C., and Poggio, T. (2001). Example-based object detection in images by components. PAMI.

[Moore et al., 1999] Moore, D. J., Essa, I. A., and Hayes Iii, M. H. (1999). Exploiting human actions and object context for recognition tasks. In ICCV.

[Nelissen et al., 2005] Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., and Orban, G. A. (2005). Observing others: multiple action representation in the frontal lobe. Science, 310:332–336.

[Niebles et al., 2008] Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. IJCV.

[Ohta et al., 1978] Ohta, Y., Kanade, T., and Sakai, T. (1978). An analysis system for scenes containing objects with substructures. In Proceedings of the Fourth International Joint Conference on Pattern Recognitions.

[Oliva and Torralba, 2001] Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV, 42:145–175.

[Oquab et al., 2014] Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In CVPR.

[Oren et al., 1997] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio, T. (1997). Pedestrian detection using wavelet templates. In CVPR.

[Palmer, 1999] Palmer, S. E. (1999). Vision science: photons to phenomenology. MIT Press, Cambridge, Mass.

[Pandey and Lazebnik, 2011] Pandey, M. and Lazebnik, S. (2011). Scene recognition and weakly supervised object localization with deformable part-based models. In ICCV.

[Papageorgiou and Poggio, 2000] Papageorgiou, C. and Poggio, T. (2000). A trainable system for object detection. IJCV.

[Payet and Todorovic, 2011] Payet, N. and Todorovic, S. (2011). Scene shape from texture of objects. In CVPR.

[Peursum et al., 2005] Peursum, P., West, G., and Venkatesh, S. (2005). Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In ICCV.

[Pishchulin et al., 2013a] Pishchulin, L., Andriluka, M., Gehler, P., and Schiele, B. (2013a). Poselet conditioned pictorial structures. In CVPR.

[Pishchulin et al., 2013b] Pishchulin, L., Andriluka, M., Gehler, P., and Schiele, B. (2013b). Strong appearance and expressive spatial models for human pose estimation. In ICCV.

[Prest et al., 2011] Prest, A., Schmid, C., and Ferrari, V. (2011). Weakly supervised learning of interactions between humans and objects. PAMI.

[Quattoni and Torralba, 2009] Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In CVPR.

[Ramanan, 2006] Ramanan, D. (2006). Learning to parse images of articulated bodies. In NIPS.

[Ramanan et al., 2005] Ramanan, D., Forsyth, D. A., and Zisserman, A. (2005). Strike a pose: Tracking people by finding stylized poses. In CVPR.

[Rodriguez et al., 2011] Rodriguez, M., Laptev, I., Sivic, J., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In ICCV.

[Rother et al., 2007] Rother, D., Patwardhan, K., and Sapiro, G. (2007). What can casual walkers tell us about the 3D scene. In CVPR.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science, 290:2323–2326.

[Russakovsky et al., 2014] Russakovsky, O., Ma, S., Krause, J., Deng, J., Berg, A., and Li, F.-F. (2014). Imagenet: Large scale visual recognition challenge 2014. http://imagenet.org/challenges/LSVRC/2014/.

[Sapp and Taskar, 2013] Sapp, B. and Taskar, B. (2013). MODEC: Multimodal decomposable models for human pose estimation. In CVPR.

[Sapp et al., 2010] Sapp, B., Toshev, A., and Taskar, B. (2010). Cascaded models for articulated pose estimation. In ECCV.

[Saptharishi et al., 2000] Saptharishi, M., Hampshire, J. B., Khosla, P. K., et al. (2000). Agent-based moving object correspondence using differential discriminative diagnosis. In CVPR.

[Satkin et al., 2012] Satkin, S., Lin, J., and Hebert, M. (2012). Data-driven scene understanding from 3D models. In Proc. BMVC.

[Schölkopf and Smola, 2002] Schölkopf, B. and Smola, A. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA.

[Schuldt et al., 2004] Schuldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: a local svm approach. In Proc. ICPR.

[Schwing et al., 2012] Schwing, A. G., Hazan, T., Pollefeys, M., and Urtasun, R. (2012). Efficient structured prediction for 3D indoor scene understanding. In CVPR.

[Shotton et al., 2006] Shotton, J., Winn, J., Rother, C., and Criminisi, A. (2006). Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In ECCV.

[Sidenbladh et al., 2000] Sidenbladh, H., Black, M. J., and Fleet, D. J. (2000). Stochastic tracking of 3d human figures using 2d image motion. In ECCV.

[Sigal et al., 2004] Sigal, L., Bhatia, S., Roth, S., Black, M. J., and Isard, M. (2004). Tracking loose-limbed people. In CVPR.

[Silberman et al., 2014] Silberman, N., Sontag, D., and Fergus, R. (2014). Instance segmentation of indoor scenes using a coverage loss. In ECCV.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In NIPS.

[Singh et al., 2012] Singh, S., Gupta, A., and Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches. In ECCV.

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In ICCV.

[Soomro et al., 2012] Soomro, K., Zamir, A. R., and Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.

[Stark et al., 2008] Stark, M., Lies, P., Zillich, M., Wyatt, J., and Schiele, B. (2008). Functional object class detection based on learned affordance cues. In ICVS.

[Staufer and Grimson, 1998] Staufer, C. and Grimson, W. (1998). Adaptive background mixture models for real-time tracking. In CVPR.

[Sun and Ponce, 2013] Sun, J. and Ponce, J. (2013). Learning discriminative part detectors for image classification and cosegmentation. In ICCV.

[Tenenbaum et al., 2000] Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2323.

[Tighe and Lazebnik, 2010] Tighe, J. and Lazebnik, S. (2010). Superparsing: scalable nonparametric image parsing with superpixels. In ECCV.

[Tompson et al., 2014] Tompson, J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In CoRR.

[Toshev and Szegedy, 2014] Toshev, A. and Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In CVPR.

[Turek et al., 2010] Turek, M., Hoogs, A., and Collins, R. (2010). Unsupervised learning of functional categories in video scenes. In ECCV.

[Uijlings et al., 2009] Uijlings, J., Smeulders, A., and Scha, R. (2009). What is the spatial extent of an object? In CVPR.

[Urgesi et al., 2006] Urgesi, C., Moro, V., Candidi, M., and Aglioti, S. M. (2006). Mapping implied body actions in the human motor system. The Journal of Neuroscience, 26:7942–7949.

[Urtasun et al., 2006] Urtasun, R., Fleet, D. J., and Fua, P. (2006). 3d people tracking with gaussian process dynamical models. In CVPR.

[Urtasun et al., 2005] Urtasun, R., Fleet, D. J., Hertzmann, A., and Fua, P. (2005). Priors for people tracking from small training setsdeutscher2000articulated. In ICCV.

[Urtasun et al., 2007] Urtasun, R., Fleet, D. J., and Lawrence, N. D. (2007). Modeling human locomotion with topologically constrained latent variable models. In ICCV Workshop on Human Motion: Understanding, Modeling, Capture and Animation.

[Vedaldi et al., 2009] Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A. (2009). Multiple kernels for object detection. In ICCV.

[Vogel and Schiele, 2004] Vogel, J. and Schiele, B. (2004). Natural scene retrieval based on a semantic modeling step, pages 207–215. Springer.

[Vu et al., 2014] Vu, T. H., Olsson, C., Oliva, A., Sivic, J., and Torralba, A. (2014). Predicting actions from static scenes. In ECCV.

[Walker et al., 2014] Walker, J., Gupta, A., and Hebert, M. (2014). Patch to the future: Unsupervised visual prediction. In CVPR.

[Wang and Li, 2013] Wang, F. and Li, Y. (2013). Beyond physical connections: Tree models in human pose estimation. In CVPR.

[Wang et al., 2010] Wang, H., Gould, S., and Koller, D. (2010). Discriminative learning with latent variables for cluttered indoor scene understanding. In ECCV.

[Wang and Schmid, 2013] Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In ICCV.

[Wang et al., 2006a] Wang, X., Tieu, K., and Grimson, E. (2006a). Learning semantic scene models by trajectory analysis. In ECCV.

[Wang et al., 2006b] Wang, Y., Jiang, H., Drew, M. S., Li, Z. N., and Mori, G. (2006b). Unsupervised discovery of action classes. In CVPR.

[Wong et al., 2007] Wong, K.-Y. K., Kim, T.-K., and Cipolla, R. (2007). Learning motion categories using both semantic and structural information. In CVPR.

[Yang et al., 2010] Yang, W., Wang, Y., and Mori, G. (2010). Recognizing human actions from still images with latent poses. In CVPR.

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

[Yang and Ramanan, 2013] Yang, Y. and Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. PAMI.

[Yao and Fei-Fei, 2010a] Yao, B. and Fei-Fei, L. (2010a). Grouplet: A structured image representation for recognizing human and object interactions. In CVPR.

[Yao and Fei-Fei, 2010b] Yao, B. and Fei-Fei, L. (2010b). Modeling mutual context of object and human pose in human-object interaction activities. In CVPR.

[Yao et al., 2011a] Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. (2011a). Human action recognition by learning bases of action attributes and parts. In ICCV.

[Yao et al., 2011b] Yao, B., Khosla, A., and Fei-Fei, L. (2011b). Combining randomization and discrimination for fine-grained image categorization. In CVPR.

[Yu and Joachims, 2009] Yu, C. and Joachims, T. (2009). Learning structural svms with latent variables. In ICML.

[Yuen and Torralba, 2010] Yuen, J. and Torralba, A. (2010). A data-driven approach for event prediction. In ECCV.

[Zhang et al., 2007] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV.

[Zhao and Zhu, 2011] Zhao, Y. and Zhu, S.-C. (2011). Image parsing with stochastic scene grammar. In NIPS.