Part Annotations via Pairwise Correspondence

Subhransu Maji **Gregory Shakhnarovich**
{smaji, greg}@ttic.edu
Toyota Technological Institute at Chicago, Chicago, IL

Abstract

We explore the use of an interface to mark pairs of points on two images which are in "correspondence" with one another, as a way of collecting part annotations. The interface allows annotations of visual categories that are structurally diverse, such as *chairs* and *buildings*, where it is difficult to define a set of parts, or landmarks, that are consistent, namable or uniquely defined across all instances of the category. It allows flexibility in annotation – the landmarks can be instance specific, are not constrained by language, could be many to one, etc and requires little category specific instructions. We compare our approach to two popular methods of collecting part annotations, (1) drawing bounding boxes for a set of parts, and (2) annotating a set of landmarks, in terms of annotation setup overhead, cost, difficulty, applicability and utility, and identify scenarios where one method is better suited than the others. Preliminary experiments suggest that such annotations between a sparse set of pairs can be used to bootstrap many high level visual recognition tasks such as part discovery and semantic saliency.

Introduction

Image annotation is commonly used today in computer vision to construct training data. The exact nature of annotation depends on the vision task at hand; when detection or segmentation is the end goal, annotators typically mark bounding boxes containing objects, or accurate object outlines. When the vision method involves local reasoning, a more detailed annotation protocol may include marking locations of keypoints or landmarks.

Keypoint annotation is labor-intensive, requires careful and possibly biasing instructions, and relies on a definition of a standardized set of keypoints. For some categories, such a set can be defined based on domain knowledge – e.g., facial and skeletal landmarks for animals. But for many categories it would be hard, perhaps impossible, to come up with keypoints that are well defined, visually distinctive and cover the object instances well. Consider for instance categories represented in Figure 1. These have one or more of the following properties:

Figure 1: Illustration of diversity and structural variation for a few categories. Rectangles of same color mark local regions corresponding across instances for each category.

Appearance variability of semantically related, or even identical, parts. Consider, for instance, the category of church buildings. While most buildings will have windows, the windows will differ dramatically in their shape and appearance across instances: rosary vs. rectangular transparent vs. stained glass etc.

Structural flexibility of a category. Different parts can appear or not in appear in different instances. There could be multiple appearances of a part, with the number of appearances varying across instances. Again consider the churches as an example. Many instances will have a spire, others will have a dome or multiple domes, or a combination of a mausoleum and a spire; the number and location of windows and doors will vary, etc.

Unnameable or unnamed landmarks. Even when we have a reasonable set of semantically defined landmarks for a category, it may not be the optimal set. We may be

missing additional landmarks that lack a standard name, but nonetheless are repeatable, detectable and informative. For buildings, this could include meeting points of architectural elements; for people, the characteristic silhouette inflection where body parts meet, etc. On the other hand, keypoints based on, say, anatomy of an animal may be impossible to accurately observe, despite meaningful definition.

Yet, as Figure 1 demonstrates, despite these difficulties one can recognize corresponding points across instances in each category, even if one would not know how to name some of them. In this paper we describe a novel approach to annotation that capitalizes on this idea. In our annotation setup, one is presented with a pair of images containing instances of objects in the category of interest, and marks pairs of keypoints that are in correspondence between the images. The nature of these keypoints and of the correspondences, and the number of corresponding pairs for a given pair of images, is left entirely to human judgment. The only input from the designers of the annotation task consists of a small set of examples of what might be considered keypoints for the category – with a clear invitation to include other types of keypoints that seem reasonable to the annotator. As we show, this approach is efficient, intuitive for the users, and flexible enough to account for structural and visual diversity in many categories.

The sparse set of annotations obtained with our interface can be used to "jump-start" a number of visual recognition tasks. In particular, we show preliminary experiments to automatically discover a library of consistent visual parts complete with an appearance detector, predict correspondences between novel images, and even reason about semantic saliency within an image.

Motivation

Our work is motivated by the success of part-based models for a variety of computer vision tasks. Typically, part-based models capture the appearance of an object category by using a separate appearance model per part, and the overall shape of the instances in the category by the configuration of part locations (Bourdev and Malik 2009; Felzenszwalb and Huttenlocher 2005; Felzenszwalb et al. 2010; Leibe, Leonardis, and Schiele 2004). This offers some flexibility in shape through deformation modeling, and some robustness to occlusions through allowing missing parts. Furthermore, sharing parts between related categories could significantly increase efficiency of learning, allowing learning new categories from very few examples (Torralba, Murphy, and Freeman 2004). Beyond object detection, parts allow for deeper understanding of visual categories: refined categorization (car model, architectural type of building, breed of an animal species), pose estimation, evaluation of similarity between object instances, etc.

Many of these models rely on careful supervised annotation in which humans indicate, for a large dataset, locations (and possibly scale and orientation) of a number of predefined landmarks/keypoints. For instance, images of animals, including people, could be annotated with anatomical features like head, shoulders, knees, facial features etc. The parts are then derived as corresponding to individual landmarks (Mori and Malik 2006) or sets of keypoints (Bourdev and Malik 2009; Bourdev et al. 2010).

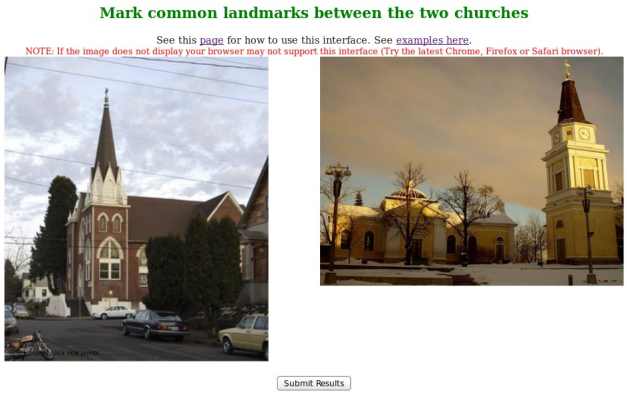


Figure 2: Interface to mark correspondences.

Approaches like these however crucially rely on the fact that all the instances of the category share the same set of keypoints. For structurally diverse categories such as chairs, sofas, airplanes or boats, any such list is likely to be incomplete or inapplicable to most instances. An important issue is that the quality of annotations suffers when such landmarks are hard to name, for e.g. corners of sofas and chairs, wingtips of airplanes etc. Significant effort is required to accurately define the semantics of these landmarks which is inherently noisy, or one has to rely on careful curation of annotations after a crowd-sourced collection. These factors severely limit the scalability of such methods to diverse categories. In this work we explore a different annotation mode, consisting of only pairwise correspondence which can address some of these issues.

Annotation setup

Interface. The annotator is presented with a pair of images of the category of interest, and asked to simply mark points in the two images that match. The interface (Figure 2) allows the user to add correspondences by first clicking on the left image and then on the corresponding point in the right image. In addition the user can adjust the locations of the clicked landmarks or delete pairs them. Once the user is done, he/she clicks a submit button.

Instructions. The user is provided with detailed instructions on how to use the interface to mark the correspondences. The pairs of images shown are generic and are *not* category specific. To guide the process of annotation we show some examples of landmarks for the category of interest such as those in Figure 3. We intentionally avoid detailed instructions or provide examples of matches, in order not to bias the users. We hope that salient landmarks will arise naturally from multiple users labeling an image. Our initial experiments validate this.

Figure 3: Example landmarks of church buildings.

Experiments. We experiment with two categories *church buildings* and *chairs* in our initial study. We collected about 300 images each of churches and chairs from Flickr and Google, which were then filtered to remove near duplicates and images containing people. These images typically contain only one prominent object, hence further annotation such as bounding boxes were not necessary. For the church buildings we collect annotations for 1000 random pairs of images, each annotated with one user on Amazon Mechanical Turk. This is roughy 2.2% of all possible pairs. For the chair category we also collected annotations for 1000 pairs. Figure 4 shows examples of such annotations.

Comparison to other annotation methods

Two popular methods that may be used to collect part annotation apart from our pairwise correspondence setup are:

Drawing part bounding boxes. Annotators are asked to draw a tight bounding box around the part of interest. This is the staple mode of annotation for rigid parts and objects such as frontal faces and pedestrians in many datasets (Everingham et al. 2010; Dalal and Triggs 2005). More recently datasets such as attributes and parts of animals (Farhadi, Endres, and Hoiem 2010) also contain bounding box annotations for parts of animals such as heads and legs, and parts of vehicles such as wheels, etc.

Marking Keypoints/Landmarks. Annotators are asked to mark the location and/or presence of a predefined set of keypoints.

There are several design choices such as cost and setup time, the importance of which depend on the application in mind and constraints on resources, etc. We compare our approach to collecting part annotations via pairwise correspondence (PC), to drawing bounding boxes (BOX) and keypoint annotation (KPT).

Annotation setup time

This includes creating instructions to the annotator, such as examples of desired output, providing clarification for ambiguous cases, etc. The PC setup time is minimal since it requires a few examples of a few images with landmarks marked on them.

Since we provide no examples of correspondences, it is worth checking if there is any agreement between annotators. Figure 5, shows the set of landmarks marked by various users on a single image for the chairs and church categories. The high level of consistency across annotators shows that we are able to get useful signal by only providing examples of generic landmarks and no detailed, category specific instructions.

Figure 5: Landmarks marked by various users on several images of chairs and church buildings. Each color denotes an annotator. The locations of landmarks provided by different annotators tend to agree at salient locations on the image.

Compared to this both BOX and KPT requires a description of the part or keypoint names, and/or examples of annotations. Often there are many ambiguities, such as left vs right, multiple parts. Our experience on collecting keypoint annotations (Maji 2011) suggests that careful instructions are necessary to avoid common annotation mistakes. The setup time greatly increases for structurally diverse categories since there are many instances where the task is ill-defined.

Annotation time

For the church category, users spend 48 seconds and marked 3 landmarks on average. For chairs, users spend 34 seconds and marked 2 landmarks on average. Note that using the pairwise annotations we get annotations for two images at the same time. As a comparison collecting keypoint annotations using the interface of (Maji 2011) takes 44 seconds and users mark 6 keypoints on average for the chair category of the PASCAL VOC 2011 dataset. Categories such as "sofa" and "aeroplane" take about 60 seconds on average. Thus our interface fares favorably in terms of the time spend by the annotators.

Annotation difficulty

When parts or landmarks are intuitive, the annotations task can be easy. One indicator of annotation difficulty is the consistency of annotations. For animal categories, the annotations for fiducial keypoints such as "eyes" and "nose" tend to be more consistent compared to the keypoints for categories such as chairs and sofas. Sometimes there are lin-

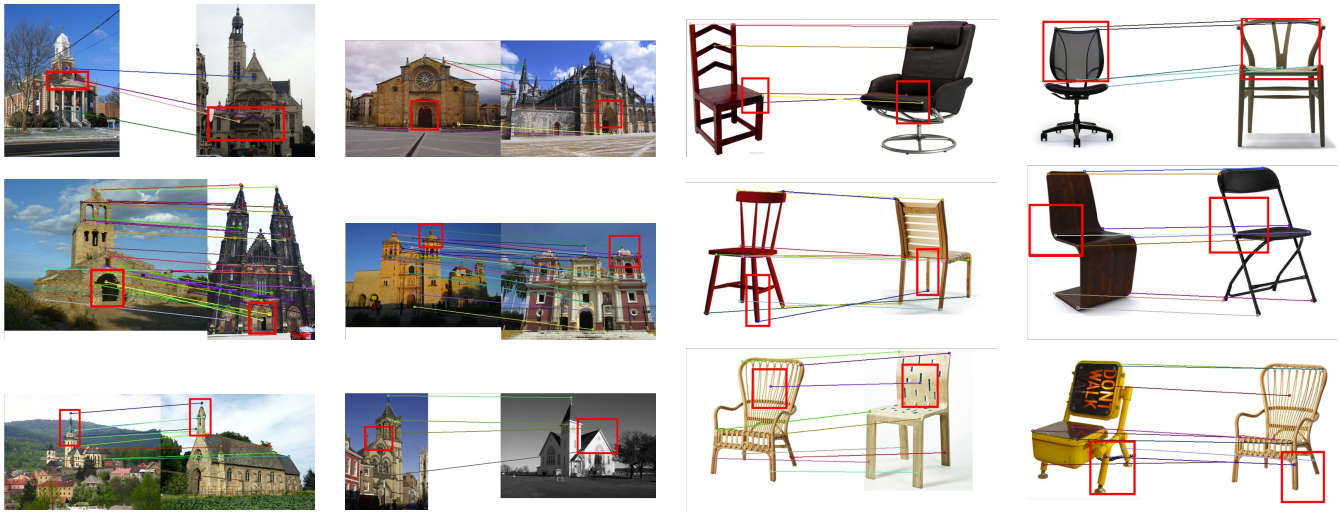


Figure 4: Some relatively good annotations collected using our interface for *church* (left) and *chair* (right) categories. Parts of one image can be matched to one another, for e.g. windows, doors and arches for churches. Given a source window in one of the images in a pair, these correspondences allow us to *automatically* find the target window in the other shown as red boxes.

guistic barriers. As an example one may not be aware of the names of parts of a horse, and there seems to a consistent confusion between the *elbow* and *knee*, even with detailed instructions.[1]

When parts don't have intuitive names, it can be difficult for annotators to remember its semantics. For "sofas", keypoints may have names such as "LEFT_BACK_TOP_CORNER" for the top right corner of the sofa's back seat or "RIGHT_FRONT_HANDLE" for the front tip of the right handle, and are often incorrectly labelled by annotators. In these cases the pairwise annotation task is much more intuitive: A landmark is defined *by* the semantic correspondence across a pair of images, in a way specific to the category at hand. By not forcing the user to adhere to a predefined set of landmarks, we avoid mistakes due to linguistic mis-interpretation or lack of careful instructions.

A difficult annotation task can significantly increase the curation time. It also increases the cost of collecting annotations, by requiring multiple redundant annotations to smooth out annotation errors.

Annotation cost

Assume for the sake of discussion that the cost of marking one bounding box in an image, marking correspondences between a pair of images and marking a set of landmarks in an image is the same.[2] If the category has a well defined set of keypoints, then for the same cost, the KP setup can be more efficient than the BOX. Consider for example the person category. Rather than separately marking bounding boxes for parts such as faces, legs, etc., the keypoint annotations allows one to create a bounding box on the fly for any combination of keypoints such as a conjunction of head and shoulders, an idea that has been used to learn parts called "poselets" (Bourdev and Malik 2009; Bourdev et al. 2010).

For these categories both KP and BOX setup are more cost effective than pairwise correspondence since each image needs to be looked only once. The pairwise correspondence setup typically would require annotating a number of pairs which is a constant factor times the number of images n. In the worst case this number could be as high as quadratic in n, but fortunately our experiments suggest that a small constant (yielding a sparse sample of edges in the image graph) may be enough for various computer vision applications.

In our experiments we sample the edges uniformly at random with replacement, but we note that better strategies for sampling edges may be possible depending on the application or prior knowledge of the distribution of images. For many large graphs, uniform random sampling has shown good empirical performance in preserving many graph properties (Leskovec and Faloutsos 2006).

Annotation utility

Pairwise annotations make the fewest assumptions on the object category, but lack the semantics of keypoints or bounding boxes, and are restricted to pairs. However the pairwise information can be propagated using the underlying semantic graph $G = (V, E)$ over the set of images. The vertex set V corresponds to images, while the edges E correspond to the collected pairwise annotations. The semantic graph can be used to derive a number of visual information which we discuss next.

Semantic saliency. Since the users are free to choose what landmarks to pick on each image, the location of the marked

[1] Horses have both knee and elbow on the front legs.

[2] The monetary cost–the amount paid to AMT workers per annotation–was in fact the same for the majority of tasks compared here.

Figure 6: Semantic saliency. Landmarks that are repeatable across other images are likely to have higher saliency (brighter intensity).

keypoints might capture a notion of semantic saliency. As we saw in Figure 5, the locations of marked landmarks on an image are highly correlated across annotators. Figure 6 shows the saliency maps for several images obtained by adding up 2D Gaussians centered at each clicked landmark in the image.

Shared parts. Given a set of corresponding landmark pairs in a pair of images, one could fit a geometric transform (in our experiments, just translation and scaling) that explain the relationship between their locations in the image. Thus, a set of landmarks that are transformed consistently might indicate a shared part. For churches, these include spires (pyramidal structure on the top of the building), windows, doors, as well as hard to name parts such as the conjunction of the roof line and the spire. Figure 4 shows several pairs of churches and chairs with the annotations and pairs of windows (shown in red) found by least square estimates of the translation and scaling.

The correspondences can be propagated by traversing the semantic graph in a breadth first manner. Thus even with sparse pairwise annotations, we can find correspondence between pairs of images as long as there is a path connecting them. Multiple occurrences of a part may be found in an image, e.g., two spires, multiple doors and windows, etc., since there could be more than one path leading to it.

Propagating pairwise correspondence accumulates noise at each step. As one can see in Figure 7(left), the similar windows found using by traversing the semantic graph are quite noisy and a appearance model learned using these examples directly may not work well. However they can be used as a rough initialization and combined with an appearance model to find better matches. Figure 7 illustrates the process using similarity based on learning a gradient histograms (Dalal and Triggs 2005) model from the source window. The model is initialized by the appearance of the source window, then updated using the top scoring matches (Figure 7(center)); final location of each match is refined by searching locally for optimal translation and scale under the current HOG model (Figure 7(right)).

Conclusion and discussion

To leverage crowd-sourcing tools such as Amazon Mechanical Turk (`http://www.mturk.com`), an important requirement is that the tasks should be intuitive for the workers. For the approach to be scalable across, say to a large number of categories, the task should also be easy to set up. Our interface for collecting pairwise correspondences tries to achieve these twin goals and can be an attractive alternative to marking a predefined set of landmarks for categories with large structural diversity. There are also disadvantages of our approach. The global semantics of the landmarks is lost, and has to be inferred from pairwise correspondences.

Our initial experiments suggest that even with sparse annotations and relying on a combination of visual similarity and the underlying semantic annotation graph, one could recover global semantics to learn parts for diverse visual categories. The computer vision aspects such a detection, description, etc, from novel images will be explored in future research.

References

[Bourdev and Malik 2009] Bourdev, L., and Malik, J. 2009. Poselets: Body part detectors trained using 3d human pose annotations. In *International Conference on Computer Vision (ICCV)*.

[Bourdev et al. 2010] Bourdev, L.; Maji, S.; Brox, T.; and Malik, J. 2010. Detecting people using mutually consistent poselet activations. In *European Conference on Computer Vision (ECCV)*.

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

[Everingham et al. 2010] Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision (IJCV)* 88(2):303–338.

[Farhadi, Endres, and Hoiem 2010] Farhadi, A.; Endres, I.; and Hoiem, D. 2010. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR)*.

[Felzenszwalb and Huttenlocher 2005] Felzenszwalb, P. F., and Huttenlocher, D. P. 2005. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)* 61:55–79.

[Felzenszwalb et al. 2010] Felzenszwalb, P.; Girshick, R.; McAllester, D.; and Ramanan, D. 2010. Object detection with discriminatively trained part-based models. *IEEE Transaction of Pattern Analysis and Machine Intelligence (PAMI)* 32(9):1627 –1645.

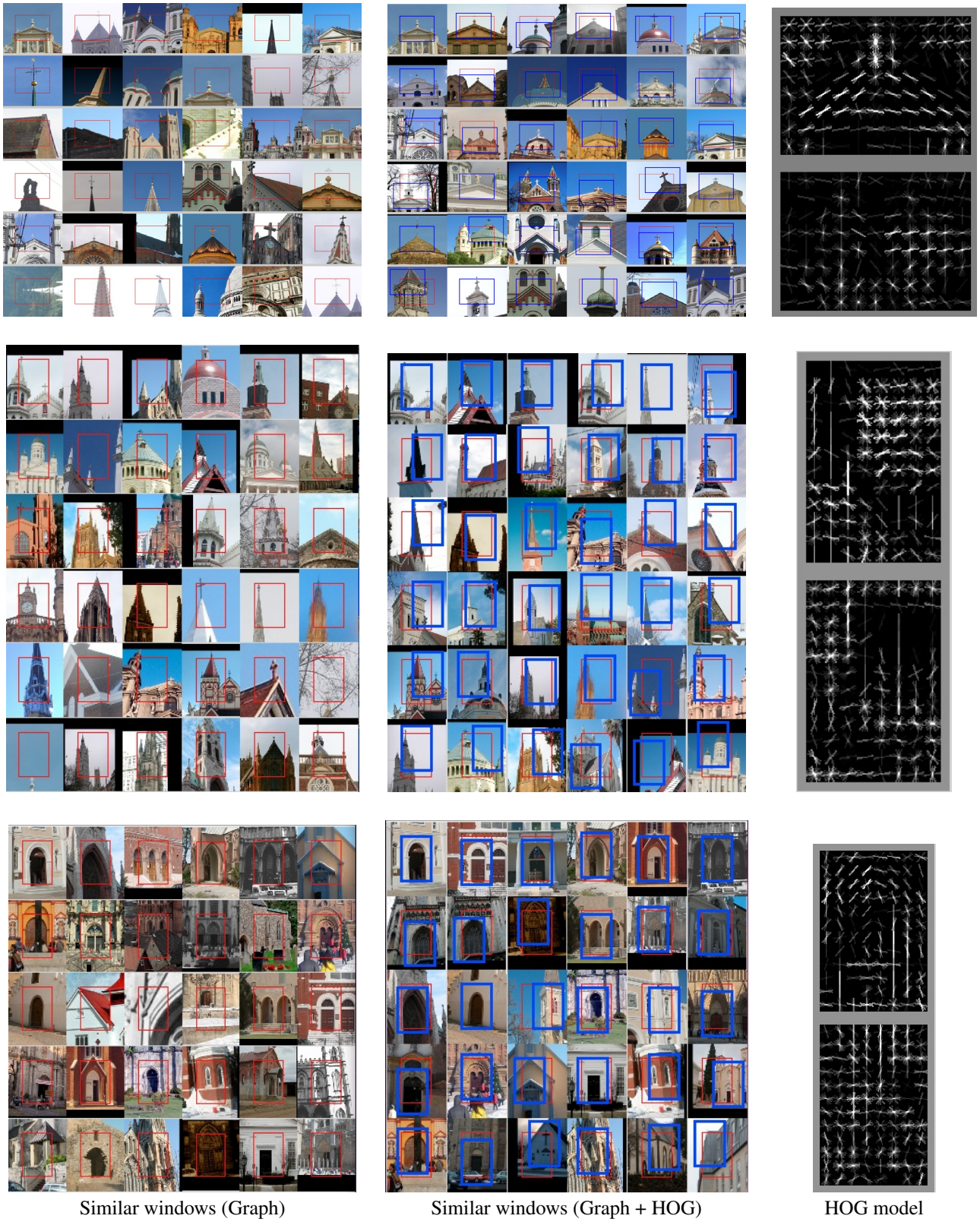
[Leibe, Leonardis, and Schiele 2004] Leibe, B.; Leonardis, A.; and Schiele, B. 2004. Combined object categorization and segmentation with an implicit shape model. In *ECCV workshop on statistical learning in computer vision*, 17–32.

[Leskovec and Faloutsos 2006] Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*.

[Maji 2011] Maji, S. 2011. Large scale image annotations on amazon mechanical turk. Technical Report UCB/EECS-2011-79, EECS Department, University of California, Berkeley.

[Mori and Malik 2006] Mori, G., and Malik, J. 2006. Recovering 3d human body configurations using shape contexts. *IEEE Transaction of Pattern Analysis and Machine Intelligence (PAMI)* 28(7).

[Torralba, Murphy, and Freeman 2004] Torralba, A.; Murphy, K.; and Freeman, W. 2004. Sharing features: efficient boosting procedures for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR)*.



Similar windows (Graph) | Similar windows (Graph + HOG) | HOG model

Figure 7: (Left) Top matches found using BFS in the annotation graph. (Center) Top matches sorted by similarity to the source window using a HOG model learned from the source window and negative images. (Right) Visualization of the coefficients of the learned HOG model: top part shows positive, bottom part negative weights. Note that the process is able to find the visually similar windows as well as refine the location (shown in blue), given the initial noisy location (shown in red).