ACM SIGGRAPH

SIS Conference Reporting

216 - Submission Reviews - By Person

Generated: August 19, 2014 04:20:19 EDT

Question	Response
Id:	papers_0216
Title:	Approximate Building Blocks for Image Analysis and Synthesis
Reviewer #91:	
1) Description	This paper proposed a new systematic approach to parsing the building blocks in the images and applied it on semantic image synthesis task. The approach is consist of three components: discriminative learning, clique detection and building block cut out. I reviewed this paper in Siggraph 2014, and this new version is greatly improved compared to Siggraph 2014 version in terms of evaluation, comparison and more exciting applications.
2) Clarity of Exposition	Clear
3) Quality of References	The references are adequate.
4) Reproducibility	The proposed approach could be reproduced. However it is very complex given so many components.
5) Rating	4.0
7) Explanation of Rating	This revision has the following significant improvements: +The extended comparison with four more methods. It shows the proposed method can detect more reasonable building blocks. The PR curve comparison is very useful addition and very informative. +The focus of the paper is much clearer with image synthesis as an application. +The evaluation and discussion is much more adequate and covered more aspect of the approach. +The new video is very nice and clearly explained the whole pipeline and each component. +The interactive editing interface is cool and the synthesize result is much improved. Some small issues: -As a systematic approach, it still suffers from incremental contribution of methodology. -Still no comparison with Wang et al. 2008, it would be better to be at least mentioned in some discussion the result section. If you can applied your method on some images in their paper, it would be very nice. Your method may not be suitable to recover the deformed and rotated blocks, thus it's doubtful if you can detect as many repeating blocks as in their method. But still, your method should be much faster and practical. -Most of the examples are 'planar', i.e. the camera is perpendicular to the building façade. There are several exceptions, but still makes me wonder how robust of the proposed method to the scale difference and distortion. It would be nice to have some evaluation or discussion on that. -The production of the video is probably rushed. There seems to be some unfinished narration sentences. Over all I think this paper is in a very good shape. I would like to see it get published.
Reviewer #3:	
1) Description	This paper presents an algorithm for finding "building blocks" inside images. These are image segments that are similar to each other and re-occur in translation. The algorithm starts by building an initial dictionary of blocks similar to [singh et al. 2012] using k-means clustering and feature re-weighting (discriminative learning), but then uses these as elements for co-occurrence analysis. The key idea is to use the distance between correspondence maps between words (patches) instead of directly the features describing these words for clustering. Next building blocks are found from these clusters using two step graph cut. This whole procedure is run multiple times combining greedily the best building blocks from different attempts based on robustness and covering (compactness).
2) Clarity of Exposition	Yes. Minor comments: Line 189: could directly BE used Line 178: restricted to images THAT HAVE a global Lines 350-358: How exactly is each cluster represented? (m_i)?
3) Quality of References	Yes
4) Reproducibility	The algorithm is somewhat involved but enough details are given. Limitations are not fully acknowledged (see below).
5) Rating	3.7
7) Explanation of Rating	This is a res-submission from Siggraph and I was a reviewer of this paper in the previous round. The authors did a good job in two aspects: first, tightening the paper and algorithm and clarifying many points (e.g. changing the "grammar" part that was not clear in the previous version by a two-step graph cut). Second, they added many missing references and comparisons to previous work. I like the idea of using the co-occurrence maps, and the examples (Figure 6) show its cleverness. I think this idea can be used in other problems where transition and near-similarity are sought. I also liked the addition of support for interactive editing (section 5.2). This applications seems very useful – allowing the user to benefit from the high level semantics defined by the "building blocks". What I don't like is that the authors still over-claim the generality of the approach (lines 276-279, 839-840). I do not think that restricting to translation is only for
Submission Information:	

Question	Response
	simplicity. I do not think that using larger groups of transformation is a trivial extension (even just in terms of the running time). Hence, the authors should tone down these claims and clearly state that the method is defined for translational repetitions of building blocks. The results are nice and extensive. I lean towards accepting the paper.

Reviewer #43:

1) Description	The paper describes a technique for parsing an image to generate a spatial cluster of repeated elements (that are approximately shifted versions of each other). It has a number of steps: unsupervised clustering based on HOG features, determination of "building blocks" through co-occurrence clustering and multi-label graph cut, and model selection via a greedy algorithm. The authors demonstrate how the decomposition using their technique can be used for image retargeting and user-assisted editing. Many examples are shown, mostly of facades.
2) Clarity of Exposition	The writing is generally clear. I'm not sure what is meant by "unreliable data" (line 83). Do the authors mean "real data with minor variations in appearance of repeated elements"? I'm not sure why the authors misspelled "co-occurrence" as "co-occurance" and "co-occurrance" in a number of places. Another typo under "Algorithm 1": "Parising".
3) Quality of References	The references are good, but the authors may want to additionally discuss: * T. Kobayashi and N. Otsu, "Image Feature Extraction Using Gradient Local Auto-Correlations," ECCV 2008 (more discriminating than HOG). * F. Schaffalitzky and A. Zisserman, "Geometric Grouping of Repeated Elements within Images" BMVC 1998 (extracts repeated elements in the scene with local and global transforms beyond mere translations). * E. Shechtman and M. Irani,"Matching Local Self-Similarities across Images and Videos," CVPR 2007 (describes a "self-similarity descriptor" at multiple scales that handles local and global geometric distortions).
4) Reproducibility	The principles are reasonably clear enough that a competent graduate student should be able to replicate the system. The authors did not show any failure examples in the paper; it would have been good to see how badly the technique can perform when the assumptions are not satisfied. While the authors show a lot of examples in the supplementary material, all of them (that I can tell) are of frontal facades. This is not good enough if the authors wish to claim any kind of generality, as the title implies. I want to see many more non-facade examples, not just the two examples in the paper (the example in Figure 1 is kind of a facade). What I am really interested is, how often does the technique fail and in such cases, how badly does it fail? If I pick an image at random from Flickr (without specifying the image category) and attempted to retarget it, what is the likelihood that it would work?
5) Rating	2.8
7) Explanation of Rating	This is generally an interesting paper; there are good technical nuggets (though their significance is somewhat reduced due to [Kalojanov et al 2012] and [Singh et al 2012]), and the results do look pretty good. However, the biggest objection I have is the assumption that the repeated elements are translated versions of each other. This limits the types of images the technique is effective on, i.e., mostly for frontal views of facades. The other objection is the selection of input images. Granted you can find many frontal views of facades and certain scenes where repeated elements happen to be approximately translations of each other; however, they constitute a tiny portion of all images available. Note that I make this comment based on the title that makes no mention of the scope of input type. There are only two non-facade examples in the paper (the example in Figure 1 is a kind of facade); all the other examples, including those in the supplementary material, are facade examples. For the topmost (sushi) example in Figure 11: what happens if the image is expanded by a length that includes a fraction of the size of the building block, say 1.5 times the horizontal length of one of the nigiri sushis instead of about 3 times? Since the authors compared the results of their technique with those of [Kwatra et al 2005], [Simakov et al 2008], [Pritch et al 2009], and [He and Sun 2012] using the examples the authors chose, it seems fair to compare the results for the images featured in these publications.

Reviewer #95:

1) Description	This paper describes a new and fully automatic method for detecting repeating elements (under translation) in an image. The method starts by non-supervised learning of discriminative features in the image, and cluster the learned features into cliques. Transformations that give rise to the instances of each block are found, and the individual instances are then segmented out using multi-label graph cuts. The description and the evaluation of this pipeline takes up the Lion's share of the paper. Next, the benefits of the detected building blocks are demonstrated using image retargeting and interactive image editing (both applications similar to those explored in the shift map editing work of Pritch et al, and the image completion work of He and Sun). I am unclear on the exact magnitude of the contribution here. Certainly image parsing sounds like a very important tool for many image processing, graphics, and vision tasks. But image parsing usually refers to assignment of a semantically meaningful labels to every pixel in the image. The current paper does not really concern itself with semantics at all, and neither it attempts to label all of the pixels in the image (produce a full segmentation). It focuses on a rather narrower task of detecting repeating elements. Furthermore, the repetitions are assumed to be related to each other by translations, so any significant rotation, scale, or perspective transformations are not handled. Although there are a few examples of other images in the paper and the supplementary materials, it seems that the method is really mostly well suited for frontal photographs of architectural facades. So my impression is that the paper's claimed contributions and scope are rather wider than the paper actually delivers.

Submission Information:

Question	Response
	Also, I agree that having identified and extracted repeating elements, and having analyzed their co-occurrence information should help with tasks such as image completion and image retargeting. And indeed the images in Figure 11 seem to indicate that the proposed technique leads to better results than existing methods. But browsing through the 600 retargeted facade images in the supplementary materials creates the impression that the performance overall is on par with that of He and Sun and Pritch et al. In some cases the proposed method produces the most realistic result of the three, but there are also many cases where there's some obvious visible artifact in its result, while one of the existing methods produces a more realistic result. (I agree that all three results are better than the pixel-based ones by Simakov et al or Kwatra et al, these pixel-based methods are indeed known to suffer from blurring artifacts).
2) Clarity of Exposition	The paper is written rather densely, and tends to be tedious to read. However, various techniques that are used in the pipeline are only briefly mentioned, even without a reference, "we train a linear support vector", "using non-maxima suppression", "Potts models with different distance metrics", "using Hungarian's algorithm", etc. Each of these may be well known, but this paper is intended for the computer graphics audience. There are multiple typos in the paper, too many to list here, it does not seem like the paper was proofread before submission.
3) Quality of References	References seem ok.
4) Reproducibility	I do not believe the paper is reproducible from the text. The pipeline is fairly involved, with multiple stages and components, each with it's own parameters, only a few of which seem to be mentioned. Limitations are not discussed as well as they should be. In particular, the limitations on the translational nature of the repetitions, and their impact of the type of images that the technique is useful for should be disclosed much more clearly and explicitly in the introduction.
5) Rating	3.0
7) Explanation of Rating	I find the proposed method interesting, and it seems to do what it does (detection) better than the existing alternatives, but I have my doubts as far as the general usefulness of the proposed approach, which I have described in the response to the first question. If the other reviewers are much more excited about this paper, I will not object to accepting it, but the paper would definitely need to be revised to provide a more balanced exposition of the scope and benefits of the contribution.

Reviewer #61:

1) Description	The authors propose an image-parsing algorithm to find building blocks (image patches) from a frontal-parallel rectified facade image. The method is unsupervised and results depend on initialization (clustering). Comparing with existing work in unsupervised image parsing in computer vision, this work is relatively narrowly focused, primarily on building façades only. For example, the transformation considered here is limited to translation alone while the current state of the art in image parsing has been relaxed to beyond linear, global transformations (allowing local, non-linear deformations). On the plus side, the authors use a SVM classifier to find discriminative features at the initial stage, which is novel and shown to be effective. The evaluation and comparison of the proposed algorithm is extensive while it remains to be clarified whether the comparison to those algorithms with a stochastic nature (results depend on random initializations) is fairly done. In general the paper addresses a computer vision problem, its two applications to computer graphics (image re-gargeting and interactive editing) are very briefly presented in this paper and lack a justification for its superior power over existing work in computer graphics. Thus this reviewer has some reservations on publishing this paper at SIGGRAPH ASIA.
2) Clarity of Exposition	The exposition is clear enough for me to understand and follow each step. But the method seems to be composed of a list of sequential steps, mostly inspired by existing work instead original. It lacks somewhat a flow of insights at a deeper level to connect all the pieces together. Figure 1 as a representative of this paper is a bit misleading, since the work presented in this paper is dealing with 2D image-parsig alone not 3D blocks. It is also limited to handle frontal-parallel building facades, no-perspective even affine deformed scenes. The GroundTruth labeling is questionable. Besides ambiguity (e.g. page 33), there seems to be some arbitrariness in Ground-Truth (GT) labeling, including the reflected versions as the same building block (e.g. page 20). One suggestion on texture synthesis results display, it is always helpful if you put the input photo together with the synthesis results. Without the input image, it is hard to judge whether your parsing and synthesis algorithms are faithful to or a reasonable variation of the input data. The text should be proofread by someone, there are typos here and there throughout the paper.
3) Quality of References	(1) Unsupervised learning of high-order structural semantics from images. In: ICCV. 2009 (2) Unsupervised detection and segmentation of identical objects. In: CVPR. (2010) Both of these papers are unsupervised methods for ‘object’ detection from a given image, using pairwise association, subgraph matching and cliques, which are directly related to current paper (3) Near-regularity Texture Analysis and Manipulation, SIGGRAPH'04.

Submission Information:

Question	Response
	This paper illustrates the limitations of MRF-type image-based texture synthesis algorithms (Efros and Leung 1999/Efros and Freeman 2001, Kwatra et al 2003) and presents a solution on how to respect input texture regularity for a class of textures called NRT, it is the first to do so. Both of the NRT texcel identification method and NRT synthesis algorithm for deformed-lattice type textures are relevant, which seems to be one of the well-motivated applications of the current paper. (4) Since reference [liu and liu] (CVPR'13: Grasp recurring patterns from a single view, short for "GRASP") is cited multiple times throughout the paper and an extensive quantitative comparison is done with authors implementation of GRASP, I am curious of how the reference to GRASP is presented in the current paper: Obviously, "GRASP" is not a symmetry detection algorithm but it is listed under 'symmetry detection' section of the related work with little explanation why. A closer look of GRASP paper suggests that it shares some parallel methodology considerations and computational framework with the proposed approach with no restrictions on transformations nor input scenes. These important similarities should be stated in the related work section upfront. The model-selection step of current work also shares some considerations stated in GRASP paper on maximizing the number of objects in a recurring pattern and the number of distinct visual words in each object.
4) Reproducibility	Yes, it is possible to reproduce the work in this paper. However, it is not clear how the current algorithm handles - missing data due to occlusion - deformations of various forms: geometry, color, shadow/lighting - homographies given its specific HOG-cell representation chosen and its expanded search space.
5) Rating	2.9
7) Explanation of Rating	The major content of this paper (8 out of 10 text pages) focuses on how to segment out a set of building blocks from a single image, using both unsupervised and supervised methods. The two applications, image retargeting and interactive editing, are relatively brief and stopped abruptly. Since image parsing, with building block discovery as a special case, is one of the central topics in computer vision, related directly to image segmentation and machine perception, SIGGRAPH may not be the most appropriate venue for this work. Even though I am quite impressed by the sheer volume of the evaluation section in this paper (and supplemental material), I also have some concerns on the fairness of the comparison. Since the proposed method has combined the results from multiple (30) runs of the authors' algorithm via a model selection process, it is not clear whether the same kind of treatment has been given to other algorithms with a stochastic nature such as GRASP etc. As for the image retargeting results evaluation, it is about time visual-inspection of large amount of synthesized images is to be replaced by more rigorous measures for retargeted image evaluation (2004 NRT paper showed one quantitative method for texture synthesis, so did Lin et al 2006 paper). Given various imperfections in the synthesized results from all three methods compared by the authors in the supplemental material, it is difficult to make a well-justified judgement on the overall output quality.

Submission Information: