Approximate Building Blocks for Image Analysis and Synthesis (papers_0216)

Author Rebuttal:

We would like to thank the reviewers for their valuable comments. We are encouraged by seeing that the reviewers agree the paper addresses an interesting problem and the results are nice and extensive. We especially appreciate reviewers who reviewed the work for the second time, and glad to see they agreed on our major improvements in evaluation and representation.

In this submission we decided to focus on image analysis and contract the discussion of the synthesis part. For people who had not seen the previous submission, this might have contributed to the perception that the paper was better suited for computer vision readers. However, the key aspect of our work is an image analysis method that extracts useful information for guiding image synthesis, which is a core computer graphics topic. This could be clarified by emphasizing how analysis and synthesis are complementary to each other, and extending the current discussion of the synthesis part for (probably) one page. We are confident that this could be done within a minor revision cycle.

Our system is designed for detecting translational repetitive patterns. In particular it is fully unsupervised, robust to appearance variation and does not rely on grid structures. As far as we know, no other methods produces better results for this task. Having additional degrees of freedom (scaling ,..., etc) may broaden the scope of applications, but are less important for our purpose of improving MRF for image retargeting/editing. In fact most of the MRF synthesizers, including Kwatra et al, Simakov et al, Pritch et al and He et al, use only translations. Additional DoF increases the search space, often leads to suboptimal solutions and computational costs. Restricting to translations is a good balance between model complexity and usability.
Our method can be extended for more general patterns. Scaling can be handled using multi-scale HoG features. Rotation and perspective is possible using local frames (Lines 840 - 843). Sever deformation/occlusion violates building block’s

definition so cannot be detected, but small ones are naturally handled by graph cut.

Our method is not restricted to a single object category (facades). This is the key benefit of being unsupervised. In contrast, the supervised sliding window detection (SSW, Figure 9) performs worse due to large appearance variations. We use the facade dataset (Zhang et al) for fair comparison. Other images, as long as they have translational patterns, will also work (c107-f107_1-a216-mat2-v1.png).

R61 and R95 suggested to have a systematic way to evaluate our retargeting results. During this rebuttal period we conducted a new user study based on submitted results: We randomly selected images from “paper0216_4_synthesis_facade\pictures folder”. Users are asked to pick the best result among three methods. In 2000 trials our method was picked for 63.9% of the time, followed by He et al (26.2%). This shows a clear statistical preference of our method. As suggested by R61 we re-run GRASP for 30 times and use our model selection to merge the results. Doing so raised GRASP’s F-score to 0.6612, still below ours 0.7007. R91 suggested additional discussion about Wang et al. 2008, where small image blocks and affine mappings are used for compression. Although Wang et al’s model is highly memory efficient, it’s not suitable for guiding retargeting/editing due to over-segmentations, as the relations between the segments may become too complicated for reasoning. We did not compare retargeting results using images featured in other publications (suggested by R43) because most of them do not have clear repetitions. Imposing unnecessary structure constraints often downgrade the results. We’ll include some examples from He et al as it aims for similar applications.

We’ll include images for some failure cases:
1) Miss-detection due to non-translational mappings: 2nd picture, top row, c107-f107_1-a216-mat2-v1.png. Some sheep are miss-detected.
2) Inaccurate detection for grids: MS, 2nd facade, page 31, c107-f107_1-a216-mat1-v1.pdf.
3) Implausible synthesis due to failure detections: paper0216_4_synthesis_facade\pictures \fac(173)_1.25_1.5.jpg. In this case He et al (middle) gives better synthesis than ours (right). This is because our detection (MS, page 57, c107-f107_1-a216-mat1-v1.pdf) did not separate

different types of windows (the blue ones and the yellow ones in GT).
4) Implausible synthesis due to complex structures: paper0216_4_synthesis_facade\pictures \fac(0)_1.5_1.5.jpg. Global relations (reflective symmetry or hierarchy) should be used.

Next we respond to individual comments. Incompatible structures (R43's "one-and-half sushi" example) is avoided by using integer steps of offset for image expansion. Graph cut further reduces conflicting offsets. "unreliable data" (Line 83) refers to noisy image data.

We agree with R61 that this paper can be balanced for graphics readers. Regarding the connection between different components, Figure 9 b) shows the improvement from each component. Discriminative learning makes detection robust to appearance variation, building block detection identifies translational patterns and the shapes of individual objects. This further improves the detection and makes the model easier to reason. Model selection avoids over-fitting. More discussions w.r.t GRASP (we use spectral clustering instead of greedy searching) will also be addressed. We'll clarify that Figure 1 is not suitable for horizontal retargeting due to perspectives. The facade dataset is the most comprehensive dataset that is currently available.

As suggested by R95, we'll explain technique terms (linear SVM ... etc). Image semantic is certainly important for recognition, but is off our topic. We will fix artifacts in the video recording as R91 suggested. Regarding to R3's question about cluster representation (Lines 350 - 358), we detect peaks from the cluster's median, and use the clique of these peaks to represent the cluster (2:26 – 2:33, c107-f107_1-a216-mat5-v1.mp4). We agree model extension is not trivial, and will re-position some related claims.