An assisted photography method for street scenes

Marynel Vázquez and Aaron Steinfeld
Robotics Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA
{marynel,steinfeld}@cmu.edu

Abstract

We present an interactive, computational approach for assisting users with visual impairments during photographic documentation of transit problems. Our technique can be described as a method to improve picture composition, while retaining visual information that is expected to be most relevant. Our system considers the position of the estimated region of interest (ROI) of a photo, and camera orientation. Saliency maps and Gestalt theory are used for guiding the user towards a more balanced picture. Our current implementation for mobile phones uses optic flow to update the internal knowledge of the position of the ROI and tilt sensor readings to correct non horizontal or vertical camera orientations. Using ground truth labels, we confirmed our method proposes valid strategies for improving image composition. Future work includes an optimized implementation and user studies.

1. Introduction

The saying, "A picture is worth a thousand words" is particularly true when documenting problems encountered in the world. Cooperation and communication between riders and transit authorities benefits riders [33]. Data collection from riders is key in this feedback loop, given that problems can be identified and documented by individuals as they move through the system. Problems related to infrastructure (e.g., stop signs), spatial conditions that prohibit movement in and around shelters, or damaged schedule signs are all examples where pictures can be effective [31]. Research suggests photos are the preferred choice of transit riders for documenting problems in public transit [30].

However, for people who are blind or low vision, taking a good picture can be problematic. A potentially effective approach is to incorporate computer vision assistance into the process. Methods for automatic image cropping [22, 32], image adaptation for small displays [7], image or video retargeting [24, 28] are possible approaches. However, these methods are designed for image post–processing, and tend to rely in composition heuristics that may not apply to photographers with visual impairments. For example, on–center compositions, where a dominant subject is geometrically centered in the image, are taken for granted in consumer photography and unlikely for users who are blind.

Figure 1. A common reporting scenario. The user roughly aims the camera at a desired target (left). If the system estimates image composition can be improved, it suggests how to move the device for a better composition (center). The photo is taken when the user reaches a good view (right).

The problem of taking a "good" picture is difficult, but dramatically simplified by the task characteristics. First, aesthetics are not an issue for problem documentation, thereby mitigating a significant challenge. Second, we do not need to know what the barrier is – we only need to know where it is. While being able to automatically annotate barriers might be useful for documentation, it is not essential. This mitigates the need for object recognition. Third, we can assume the rider is able to localize the barrier in space and roughly aim a camera at the target. This means only small camera motions are needed to balance photo composition and correct unwanted camera orientation.

Our main contribution is integrating user interaction during the image capturing process, such that users can take better pictures in real time (Figure 1). Our approach can be described as a method to avoid leaving out information that is expected to be most relevant. Motivated by Gestalt the-

978-1-4244-9497-2/10/$26.00 ©2010 IEEE

ories [10], this work evaluates a meaningful group of contiguous salient points in a picture as potential region of interest (ROI). Our system tries to center this region in a photo and correct excessive roll. Centering can benefit other vision tasks, such as those related to image retargeting. Image encoding standards, as the JPEG2000, can further exploit the region of interest by coding and transmitting it with better quality and less distortion than the rest of the image [8].

Our system automatically suggests where to move the camera. Similar to the recent project of Bae *et al*. [2], we employ computer vision techniques to estimate the camera motion required to reach a desired view. Our initial implementation for mobile devices uses optic flow to track the region of interest with respect to its initial position in the image. Measurements from accelerometers, already common in these devices, are proposed as a cheap processing alternative to detect and reduce excessive roll. Non horizontal or vertical camera orientations for street–level photographs are not desired due to their potential to confuse third party understanding of documented problems.

An optimized implementation for real–time interaction still needs to be attained, since our initial motion estimation approach is only able to cope with little or no parallax during slow camera motions. The problems inherit in processing dynamic scenes are left as future work. User studies are also planned for a complete evaluation of our method.

2. Background

Autonomous camera control systems are popular in robotics, where motion commands are less noisy than human actions. Dixon *et al*. [12], for example, describe the implementation of a robot photographer aimed at capturing "good" pictures of people. Desnoyer and Wettergreen [9] work towards aesthetically aware autonomous agents.

The interest generated by pictures in third–party observers has been studied in consumer photography [27]. Experimental results indicate interest is driven by influences of people, composition and subject matters. Luo *et al*. [20] posit that a good composition is the most influential attribute for image emphasis selection. Their system evaluates composition using saliency maps, and estimates which image receives the most attention from a set of an event.

Salient regions in images tend to be considered as information carriers that deliver the photographer's intention, and catches part of the observers' attention as a whole [7]. In our application domain, stimulus–driven visual attention is good cue for finding the regions of interest in street pictures, since transit elements tend to be salient and of high contrast. The number, type and combination strategy of features used for detecting saliency in images is a problem of its own. Different situations call for different solutions, and so we chose to test several methods with a variety of street photographs (see Section 4). Other authors have further demonstrated the benefit of including cognitive factors such as knowledge, expectations and current goals in image understanding processes. This extension is left as future work for our application. Interested readers in the subject are encouraged to read [14].

Strategies for prominent region selection tend to be strongly biased towards the point of maximum saliency. Rutishauser *et al*. [26] and Siagian and Itti [29] used a region growing algorithm and adaptive thresholding to segment regions given the most salient location. Frintrop [13] computed focus of attention in images by connecting pixels that differ at most 25% from the maximum, as suggested experimentally. Walther and Koch [35] based their thresholding procedure in a neural network of linear threshold units, and labeled points around maximum saliency using a connected components algorithm. Unfortunately, these approaches are not suited to our documentation problem. Figure 2 shows an example where a window of the building in the background has maximum saliency, even though no window is expected to be more important than the bus sign.

Ma and Zhang's [21] saliency segmentation method shares the same spirit of our approach, since they consider principles of Gestalt theory [19] to fit rectangles to attended regions in contrast maps. Their method consists in looking for the optimal partition of attended and unattended fuzzy areas in these maps, where attended contrast points with a value greater than a threshold serve as seeds for fuzzy growing. Wang and Li [36] also grow attended regions based on similarity and proximity Gestalt principles. First, the most representative block of salient pixels is detected in the largest saliency component of an image. Then, the block is extended by looking at similar neighbors. Unlike these methods, we use Gestalt theory to approach both the partition and selection of relevant attended areas. Speed is crucial for our real time application, and simple, approximate solutions are valuable for its mobile implementation.

Deville *et al*. [11] developed an alerting system to attract the attention of people with visual impairments to regions of interest. The authors describe a mobility aid for blind users, based on visual substitution by the auditory channel. Depth gradient from stereoscopic cameras and color are proposed to detect salient regions in images of the surrounding. Sounds are used to indicate noteworthy parts of the scene, which suggests auditory feedback for our system.

Figure 2. Initial view of a street scene (left), corresponding saliency map (center), and view proposed by our system (right).

3. Assisted Photography Method

Our method for improving image composition is based on estimating a relative motion of the camera that would result in a "better" view of the scene. Our procedure starts by estimating saliency to describe image composition. Then, quality assessment is performed according to how well this image fits our attention–getting composition model. If there is evidence that suggests the picture can be improved, a change in camera view is proposed and the user is guided towards taking a better photo. Image improvement is achieved through small camera motions intended to center the estimated region of interest, and fix unwanted orientation (i.e. excessive roll). Centering can help document problems as it drives composition towards highlighting evidence in the middle of pictures, and increases the chance that the surrounding content will include relevant context.

The selection of the region of interest depends on saliency estimation, and the importance of image parts changes according to the arrangement of elements in pictures. Therefore, we only estimate the location of the ROI from saliency when its position cannot be inferred from past estimations. Selecting a new ROI every time the camera moves can easily generate confusing motions for image improvement. Suppose you estimate saliency in Figure 2, and select as ROI a region that encloses most of the bus sign. Now imagine that a yellow, high–contrast element becomes part of the picture after moving a little to center the sign. Re–estimating the ROI from current saliency could direct the user towards centering the yellow element, instead of the sign as before. This could lead to large shifts away from the object the user is attempting to document.

We estimate global motion from optic flow every time a new picture is captured, and use this information to predict the position of the ROI. We use the popular Lucas–Kanade feature tracker [5] in an Expectation-Maximization framework, which allows to detect outlier motion vectors. FAST features [23] are selected every two frames (akin to [34]), which eases the problems of adding and discarding specific features along image sequences. An affine model is finally used to describe motion, though we desire a more robust approach. Methods related to augmented reality [18] are valid for improving our system. We expect camera motion estimation to aid in situations where fast and abrupt device control aggravates image understanding processes.

3.1. Image composition assessment

We model the target of a photo as its most meaningful group of contiguous salient regions. Our strategy was designed for the transit domain without explicit knowledge of object models. This leverages the fact that this domain is strongly composed by conspicuous elements.

3.1.1 Selecting the region of interest

Our initial implementation constructs a model of visual attention in an image employing a simplified version of saliency maps, as defined by Itti and Koch [17]. Our image features are intensity, and (red–green and blue–yellow) color opponency. Our normalization operator for fusing feature maps follows Frintrop's *uniqueness weight function* definition [13]. Our experimental results support this approach, though it was selected for convenience at first. Other saliency estimation methods could be used if desired.

Once a saliency map S from a picture is constructed, we select as ROI its most relevant group of contiguous salient points. Consider a discretized version $\hat{S}$ of S as a 2D histogram, and suppose samples are uniformly and independently distributed along all bins. We want to find an approximation of the most meaningful contiguous group of bins where a high, unexpected, amount of samples were placed.

The probability of a sample falling into a region $B = \{b_1, \ldots, b_n\}$ of n bins is $p(B) = n/L$, where $L = wh$ is the total number of bins. On the other hand, the density r of this region is $r(B) = k(B)/M$, where $k(B) = \sum_{i=1}^{n} \hat{S}(b_i)$ is the number of samples that fall in B, and $M = \sum_{x,y} \hat{S}(x, y)$ is the total number of samples in $\hat{S}$.

We then consider the relative entropy of a region B (assuming a prior uniform distribution) as

$$H(B) = \begin{cases} 0, \text{ if } r(B) \leq p(B) \\ r(B) \log \frac{r(B)}{p(B)} + (1 - r(B)) \log \frac{1-r(B)}{1-p(B)}, \text{ otherwise} \end{cases}$$

A *meaningful interesting region* is one that satisfies $H(B) > (1/M) \log(L(L+1)/2)$. As explained by Desolneux *et al*. [10], regions are more meaningful as H increases.

The region with maximum relative entropy is our desired ROI. A complete search for this region requires adapting to the distribution of samples to increase H. In practice, however, we threshold the discretized saliency map by

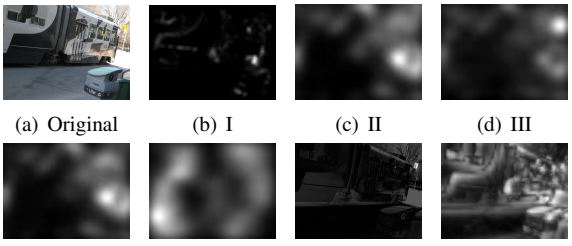
$$t = \frac{\sum_{x,y} \hat{S}(x, y)}{L} \tag{1}$$

and consider groups of contiguous bins that posses at least t samples as potential meaningful interesting regions. Using Chang *et al*. [6] linear–time component–labeling algorithm, we connect bins in the thresholded map. The ROI is finally chosen by comparing H for the connected components.

3.1.2 Image quality

The weighted mean of the selected ROI is the image center proposed by our system, as shown in Figure 3. The saliency map $\hat{S}$ is used to weight the points belonging to the ROI. If the weighted mean is near the geometric center of the picture, then our system considers the target to be well placed in the image for problem documentation. In this case, the system only tries to correct for excessive roll.

Figure 3. Left to right: saliency; region of interest (in white); and max. saliency (circle) and weighted mean of the ROI (rectangle).



(a) Original (b) I (c) II (d) III

(e) IV(a) (f) IV(b) (g) V(a) (h) VI

Figure 4. Saliency maps (methods I, . . . , VI) can vary significantly for a single image. See Section 4.1 for more details.

We assume horizontal or vertical camera orientations are optimal in the Manhattan world of street photographs. We rely on tilt sensors, now common in camera phones, for estimating the orientation of the device.

Special preference is given to centering the region of interest before correcting orientation. Framing the ROI is the priority during problem documentation. The direction of the motion proposed to the user is given by the vector from the middle of images to the weighted mean of the ROI. The rotation is the opposite of the extra roll of the camera.

The size of the region of interest is considered indirectly by our method. There is a trade–off between the size of a the region and its meaningfulness. If a region is too small, other salient regions have higher probability of being preferred. If it is too big, it might end up being not meaningful at all.

4. Experiments and Results

We studied the behavior of our system for various configurations, and compared it with a technique for automatic thumbnail cropping. Our data set consists of 776 street photos from the team and *LabelMe* [25].

4.1. Saliency and ROI estimation

We conducted a labeling exercise to evaluate the capacity of our algorithm for proposing new image centers. The images from our data set have a length of 1280px, and are generally aligned with respect to the horizon or vertical structures. To reduce their quality, we randomly selected a region (tile) with a length of 640px inside them. This tile was allowed to be rotated by an angle between $-15°$ and $15°$. Three people then manually labeled where the center of these tiles should be located, based on balancing composition and without looking at the corresponding full image.

The mean distance between the labeled centers collected per tile was close to 100px, and so we chose this value for filtering the labeled centers used in the test, and evaluating the performance of re-centering approaches. The labeled centers placed near the middle of the image (at most 100px from the middle) were ignored, and those in the periphery (at least 100px away) became our ground truth. The filtering led to a total of 607 images used for comparison.

A re-centering method succeeds if its proposed image center is less than 100px from any of the ground truth centers. If the desired center is far from all ground truth points, then the approach fails. In this case we lack evidence that supports the selected center as a valid point towards which guide the user. Note this evaluation is subjective to labelers' opinions, and we do not have a method of telling which labeler is right when there is disagreement.

We generated saliency maps for the tiles with several methods, because there is no approach that works best in all circumstances (Fig. 4). We considered,

I. Walther and Koch's biologically inspired model [35].
II. Hou and Zhang's spectral residual approach [16].
III. Guo *et al*. method [15] for color images.
IV. Bian and Zhang's spectral domain approach [3] for grayscale images (a), and for YUV color images (b).
V. Achanta *et al*. method [1] with defined boundaries (a), and with smoothing (b).
VI. Our saliency estimation method as in Section 3.1.1.

Once saliency is computed, we estimated the regions of interest in the tiles using thumbnail cropping by Suh *et al*. [32], and our technique as presented in Section 3.1.1. The work of Suh *et al*. is relevant because their aim is to estimate the most informative part of an image. Their method consists in adaptively thresholding saliency, and then fitting a rectangle to the remaining salient regions. Given our application domain, we did not prioritize saliency near the middle of the tiles, nor used face detection as in their paper.

Table 1 presents the proportion of times an approach succeeded in proposing a valid image center in the periphery. The results for selecting as new center the weighted mean of the thumbnails (using saliency for the weights) were better than just picking the center of the rectangle. The rectangles selected by the thumbnail approach were big with respect to the size of the tiles, hence the center of the rectangles is highly biased towards the middle of the images.

Figure 5 shows some results of our method, which generally did better than the thumbnail approach. We obtained poor performance with saliency (V), due to the well defined boundaries that characterize saliency maps by Achanta *et*

al. The boundaries generate an over–segmentation effect that lowers the relative entropy of potential ROI.

If we do not filter labeled centers to enforce evaluating the periphery, then both methods tend to perform well, with an agreement over 70%. Selecting the center of the thumbnails gives best results in this case, because of the biasing mentioned above. Note, however, that this evaluation is less interesting for our application. Selecting the middle of the images as new center for image improvement is equivalent to only correcting for unwanted camera orientation.

4.2. Saliency threshold

We tested the effect of varying the threshold t, as defined in equation (1), on our ROI estimation procedure. Values of $1.0t$, $1.25t$, $1.5t$ and $1.75t$ were chosen for thresholding images and tiles from our data set.

There appears to be an inverse correlation between the threshold set for saliency segmentation and relative entropy (Fig. 6). This suggests that setting a higher threshold is not appropriate since the chosen region becomes less meaningful when focusing on just a small salient section of the picture. In addition, there appears to be a positive correlation between threshold and distance from the weighted mean of the ROI to the middle of pictures. This can be explained by the bias towards the maximum salient point in the image as the threshold grows, versus a less aggressive region of interest estimation with lower thresholds.

Saliency Method	**Thumbnail Cropping** [32] (center)	(wmean)	**Our approach** (wmean)
I	0.19	0.28	0.32
II	0.23	0.33	0.37
III	0.22	0.31	0.38
IV(a)	0.24	0.34	0.37
IV(b)	0.18	0.22	0.21
V(a)	0.21	0.28	0.23
V(b)	0.20	0.28	0.26
VI	0.22	0.30	0.41

Table 1. Automatically–proposed image center evaluation for composition improvement. The higher the value, the more a method agreed with ground truth data (1.0 is total agreement).

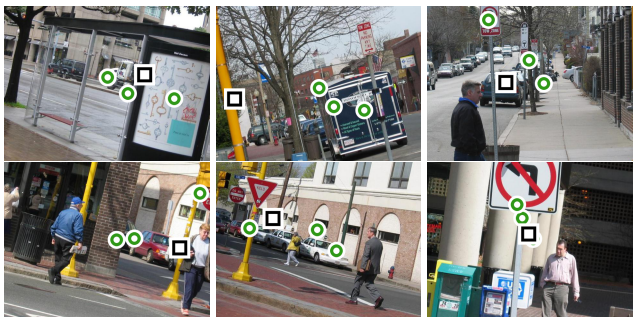


Figure 5. Labeled centers (circles) and image center suggested automatically by our method (square).

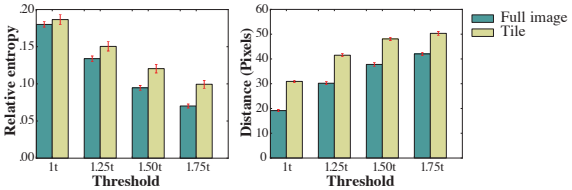


Figure 6. Mean relative entropy, and distance from the ROI to the middle of images for different saliency thresholds.

When using a bounding box to enclose the region with maximum relative entropy (Fig. 7), the mean area of the box for a threshold of $1.0t$ was 77% the size of the full pictures, with standard deviation σ of 0.18. For the tiles, the mean area of this box was 67% ($\sigma = 0.25$). This suggests that big interesting regions are considered by our procedure to describe visual attentive areas. Still, our strategy tends to favor more specific regions in closer views of the scene. This behavior seems appropriate for estimating the ROI in images without knowing about their semantics. More information about the context in street scenes is captured with a wider view and, therefore, a more diverse visual stimuli is expected. One may argue that a lower threshold than t should be used because a more accurate approximation to the most meaningful interesting region can be found. Certainly more comparisons need to be performed, though the limited processing capabilities of mobile devices may restrict our implementation to approximate solutions.

5. Conclusions and Future Work

We have presented a framework for assisted photography aimed at helping riders with visual impairments during transit problem documentation. Unlike other approaches for mobile accessibility that ask for human assistance in order to complete a task (e.g. finding specific objects [4]), we are developing an interactive application that provides fast, real–time user feedback. Our current implementation for the iPhone platform relies on saliency estimation, Gestalt theory and optic flow for guiding the user towards a better view of the scene. The device's tilt sensor is also leveraged

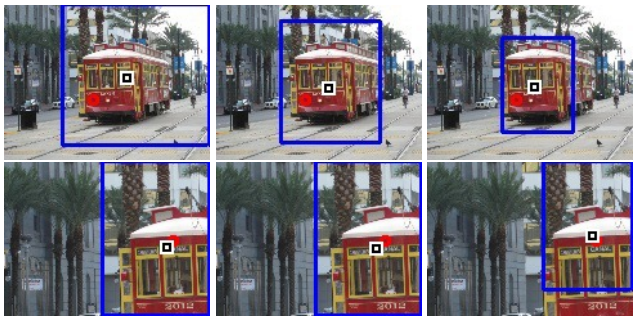


Figure 7. Enclosed ROI in full image (top row) and tiles (bottom) using thresholds of $1t$,$1.25t$ and $1.5t$. Red circle denotes max. saliency. Black and white rectangle is weighted mean of the ROI.

to detect and correct unwanted camera orientations without the need for extra high computational power. Empirical evidence suggests composition improvement can be achieved by estimating meaningful salient regions in images.

We believe this step towards on-site assisted photography sets strong foundations for assisted photography during problem documentation. We expect the incorporation of semantic information into our framework to improve results. Planned future work includes user tests for a realistic evaluation of the method. We still need to account for dynamic scenes and their added complexity.

6. Acknowledgments

This work was funded by grant number H133E080019 from the United States Department of Education through the National Institute on Disability and Rehabilitation Research. We thank Alexei Efros and Martial Hebert for their valuable comments.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Proc. CVPR*, pages 1597–1604, 2009.

[2] S. Bae, A. Agarwala, and F. Durand. Computational rephotography. *ACM Trans. Graph.*, 29(3):1–15, 2010.

[3] P. Bian and L. Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *Proc. ICONIP*, pages 251–258, 2008.

[4] J. P. Bigham, C. Jayant, A. Miller, B. White, and T. Yeh. VizWiz::LocateIt - Enabeling Blind People to Locate Objects in Their Environment. In *Proc. CVAVI*, 2010.

[5] J.-Y. Bouget. Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the algorithm. Technical report, MRL, Intel Corporation, 1999.

[6] F. Chang, C.-J. Chen, and C.-J. Lu. A linear-time component-labeling algorithm using contour tracing technique. *Comput. Vis. Image Underst.*, 93(2):206–220, 2004.

[7] L. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H. Zhou. A visual attention model for adapting images on small displays. Technical Report MSR-TR-2002-125, Microsoft Research, 2002.

[8] C. Christopoulos, A. Skodras, and T. Ebrahimi. The JPEG2000 still image coding system: an overview. *IEEE Trans. Consum. Electron.*, 46(4):1103–1127, Nov. 2000.

[9] M. Desnoyer and D. Wettergreen. Aesthetic Image Classification for Autonomous Agents. In *Proc. ICPR*, 2010.

[10] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. 2007.

[11] B. Deville, G. Bologna, M. Vinckenbosch, and T. Pun. Guiding the focus of attention of blind people with visual saliency. In *Proc. CVAVI*, 2008.

[12] M. Dixon, C. M. Grimm, and W. D. Smart. Picture composition for a robot photographer. Technical Report WUCSE-2003-52, Washington University in St. Louis, 2003.

[13] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. PhD thesis, University of Bonn, 2006.

[14] S. Frintrop, E. Rome, and H. I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7(1):1–39, 2010.

[15] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. CVPR*, 2008.

[16] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. CVPR*, 2007.

[17] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[18] G. Klein and D. Murray. Parallel tracking and mapping on a camera phone. In *Proc. ISMAR*, pages 83–86, 2009.

[19] K. Koffka. *Principles of Gestalt Psychology*. Harcourt, Brace & Company, 1935.

[20] J. Luo, A. Singhal, and A. Savakis. Efficient mobile imaging using emphasis image selection. In *Proc. PICS*, 2003.

[21] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *Proc. ACM MM*, 2003.

[22] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Proc. ICCV*, 2009.

[23] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, pages 430–443. 2006.

[24] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Trans. Graph.*, 27(3):1–9, 2008.

[25] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[26] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition? In *Proc. CVPR*, 2004.

[27] A. E. Savakis, S. P. Etz, and A. C. Loui. Evaluation of image appeal in consumer photography. In *Proc. SPIE*, 2000.

[28] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch. Automatic image retargeting. In *Proc. MUM*, 2005.

[29] C. Siagian and L. Itti. Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, 25(4):861–873, July 2009.

[30] A. Steinfeld, R. Aziz, L. Von Dehsen, S. Y. Park, J. Maisel, and E. Steinfeld. Modality preference for rider reports on transit accessibility problems. TRB 2010 Annual Meeting. Washington, DC: Transportation Research Board, 2010.

[31] A. Steinfeld, J. Maisel, and E. Steinfeld. The value of citizen science to promote transit accessibility. In *First International Symposium on Quality of Life Technology*, July 2009.

[32] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proc. UIST*, 2003.

[33] M. M. Uslan, A. F. Peck, W. R. Wiener, and A. Stern. *Access to mass transit for blind and visually impaired travelers*. AFB Press, 1990.

[34] M. Vazquez and C. Chang. Real–time video smoothing for small RC helicopters. In *Proc. SMC*, 2009.

[35] D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006.

[36] Z. Wang and B. Li. A two-stage approach to saliency detection in images. In *Proc. ICASSP*, pages 965 –968, mar. 2008.