1

AVA: A Large-Scale Database for Aesthetic Visual Analysis

Wei-Ta Chu

National Chung Cheng University

N. Murray, L. Marchesotti, and F. Perronnin, “AVA: A Large-Scale Database for Aesthetic Visual Analysis,” CVPR, 2012.

Introduction

- Novel datasets shared by the community will greatly advance the aesthetic visual analysis research.
- To date, at most 20,000 images have been used to train aesthetic models.
- Contributions:
  - Introduce a novel large-scale database (250,000 images)
  - Explore the factors that make this problem challenging
  - Show that not only does the *scale* of training data matter for increasing performance, but also the *aesthetic quality* of the images for training.

Creating AVA

- Collect images from www.dpchallenge.com
- In the community, images are uploaded and scored in response to photographic challenges.
- Create AVA by collecting approximately 255,000 images covering a wide variety of subjects on 1,447 challenges. After combination, it reduces to 963 challenges. Each image is associated with a single challenge.

Figure 1. A sample challenge entitled "Skyscape" from the social network www.dpchallenge.com. Images are ranked according to average score and the top three are awarded ribbons.

Creating AVA

- Aesthetic annotations
  - Each image is associated with a distribution of scores which correspond to individual votes.
  - The number of votes per image ranges from 78 to 549, with an average of 210 votes.
- Semantic annotations
  - 66 textual tags describing the semantics of images
  - Approximately 200,000 images contain at least one tag, and 150,000 images contain 2 tags.

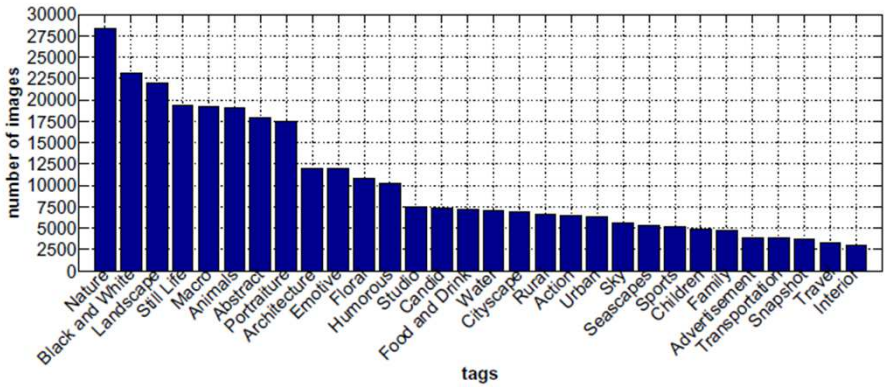

Figure 2. Frequency of the 30 most common semantic tags in AVA.

Creating AVA

- Photographic style annotations
  - Manually select 72 challenges corresponding to photographic styles and identify three broad categories according to a popular photography manual: *Light*, *Colour*, *Composition*.
  - 14 photographic styles along with the number of associated images: Complementary colors (949), Duotones (1,301), High dynamic range (396), Image grain (840), Light on white (1,199), Long exposure (845), Macro (1,698), Motion blur (609), Negative image (959), Rule of thirds (1,031), Shallow DOF (710), Silhouettes (1,389), Soft focus (1,479), Vanishing point (674)

AVA and Related Databases

	AVA	Photo.net	CUHK	CUHKPQ	ImageCLEF
Large Scale	Y	N	N	N	N
Scores distr.	Y	Y	N	N	N
Rich annotations	Y	N	Y	Y	Y
Semantic Labels	Y	N	N	Y	Y
Style Labels	Y	N	N	N	Y

Table 1. Comparison of the properties of current databases containing aesthetic annotations. AVA is large-scale and contains score distributions, rich annotations, and semantic and style labels.

- PN: 3,581 images. Scores 1~7. Bias problem.
- CUHK: 12,000 images. Half high quality, half low quality. Contain images with a very clear consensus on their score.
- CUHKPQ: 17,613 images. Either high or low quality.
- ImageCLEF: Lacks rich aesthetic preference annotation. Only the “interestingness” flag is available.

Analysis of AVA

- **Score distributions are largely Gaussian.**

Mean score	Average RMSE		
	Gaussian	Γ	Γ′
1-2	0.1138	**0.0717**	0.1249
2-3	0.0579	**0.0460**	0.0633
3-4	**0.0279**	0.0444	0.0325
4-5	**0.0291**	0.0412	0.0389
5-6	**0.0288**	0.0321	0.0445
6-7	0.0260	**0.0250**	0.0455
7-8	**0.0268**	0.0273	0.0424
8-9	0.0532	0.0591	**0.0403**
Average RMSE	**0.0284**	0.0335	0.0429

Table 2. Goodness-of-Fit per distribution with respect to mean score: The last row shows the average RMSE for all images in the dataset. The Gaussian distribution was the best-performing model for 62% of images in AVA.

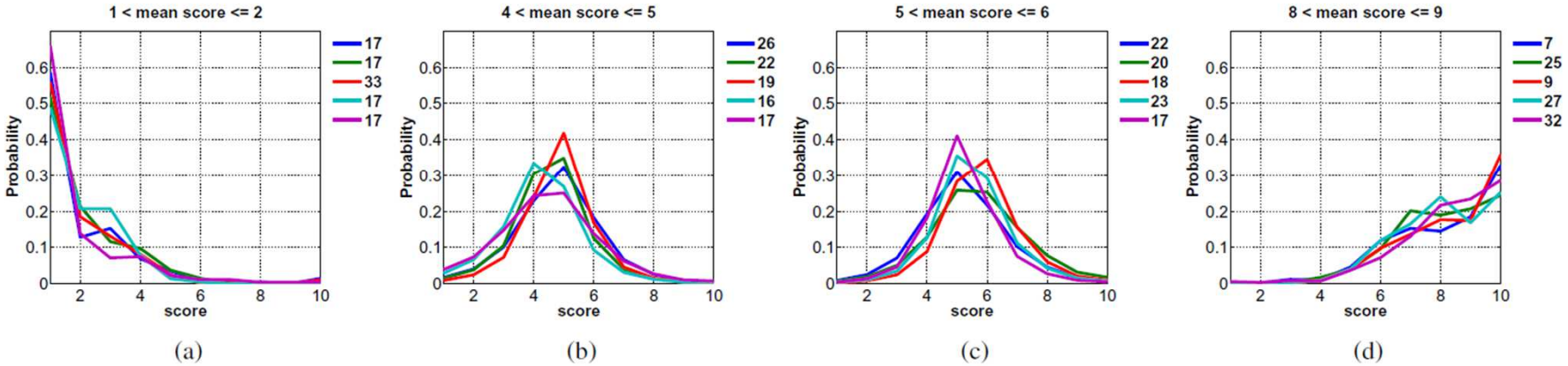

Figure 3. Clusters of distributions for images with different mean scores. The legend of each plot shows the percentage of these images associated with each cluster. Distributions with mean scores close to the mid-point of the rating scale tend to be Gaussian, with highly-skewed distributions appearing at the end-points of the scale.

Analysis of AVA

- **Standard deviation is a function of mean score.**
  - Images with “average” scores (scores around 4, 5, and 6) tend to have a lower variance than images with scores greater than 6.5 or less than 4.5.

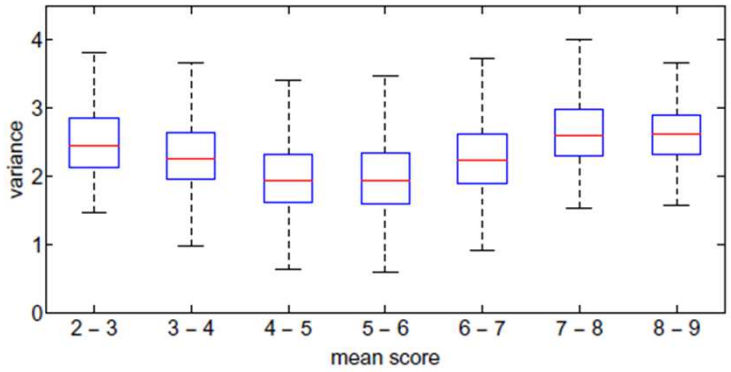

Figure 4. Distributions of variances of score distributions, for images with different mean scores. The variance tends to increase with the distance between the mean score and the mid-point of the rating scale.

Analysis of AVA

- **Images with high variance are often non-conventional.**
  - For a given mean value, images with a high variance seem more likely to be edgy or subject to interpretation.

		variance	
		low	high
mean	low	poor, conventional technique and/or subject matter	poor, non-conventional technique and/or subject matter
	high	good, conventional technique and/or subject matter	good, non-conventional technique and/or subject matter

Table 3. Mean-variance matrix. Images can be roughly divided into 4 quadrants according to conventionality and quality.

Figure 5. Examples of images with mean scores around 5 but with different score variances. High-variance images have non-conventional styles or subjects.

Semantic Content and Aesthetic Preference

- The aggregated statistics for each challenge using the score distributions of the images.
- Two “master’s students” (where only members who have won awards in previous challenges are allowed to participate) were among the top 5 scoring challenges.
- In the lowest-scoring challenges, photographers were instructed to depict or interpret the emotion or concept of the challenge’s title. This biases the aesthetic judgments towards smaller scores.

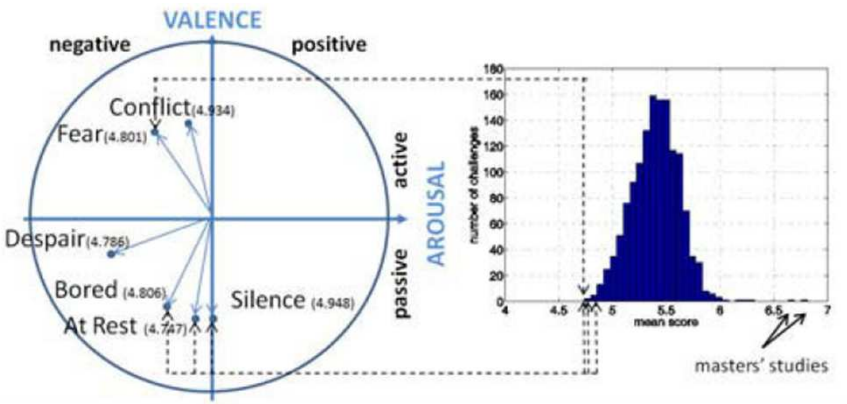

Figure 6. Challenges with a lower-than-normal average vote are often in the left quadrants of the arousal-valence plane. The two outliers on the right are masters’ studies challenges.

Semantic Content and Aesthetic Preference

- The majority of free study challenges were among the bottom 100 challenges by variance, with 11 free studies among the bottom 20 challenges.
- Challenges with specific requirements tend to lead to a greater variance of opinion.

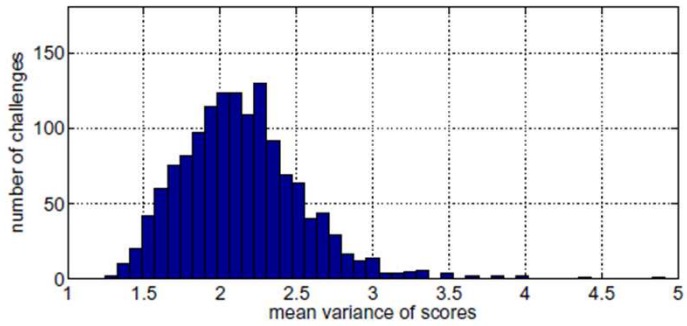

Figure 7. Histogram of the mean variance of score distributions over all challenges. Free studies tend to have low-variance score distributions.

Large-Scale Aesthetic Quality Categorization

- Treat aesthetic visual analysis as a regression problem.
- We trained linear SVMs with Stochastic Gradient Descent (SGD) on Fisher Vector (FV) signatures computed from color and SIFT descriptors .
- **The scale matters.** We consistently increase the performance with more training images.

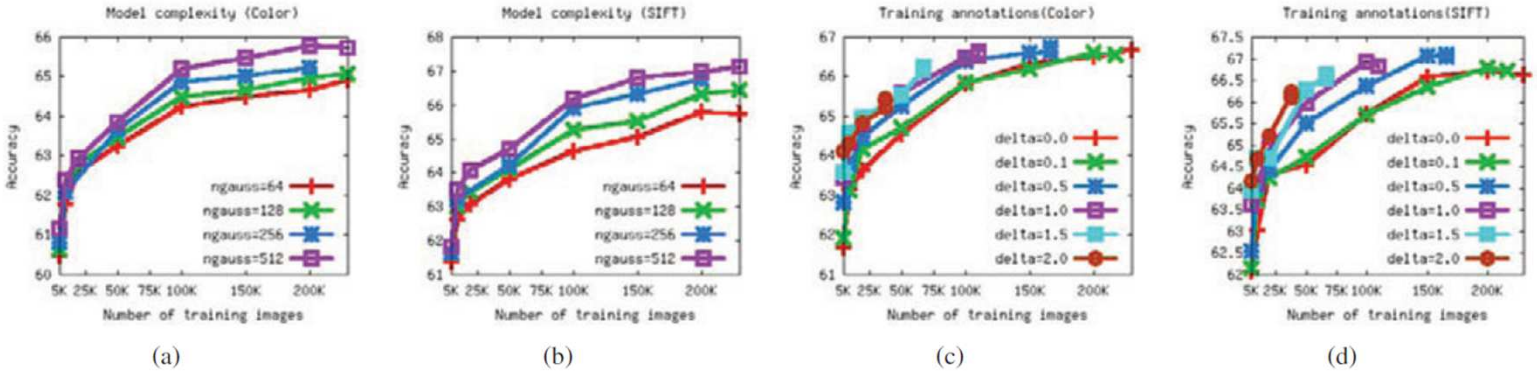

Figure 8. Results for large-scale aesthetic quality categorization for increasing model complexity ((a) and (b)) and increasing values of δ ((c) and (d)).

L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. ICCV, 2011

Large-Scale Aesthetic Quality Categorization

- **The type of training images matters.**
  - We discard from the training set all those images with an average score between $5-\delta$ and $5+\delta$.
  - For the same number of training images, the accuracy increase with δ
  - The same level of accuracy achieved by increasing training samples can also be achieved by increasing δ

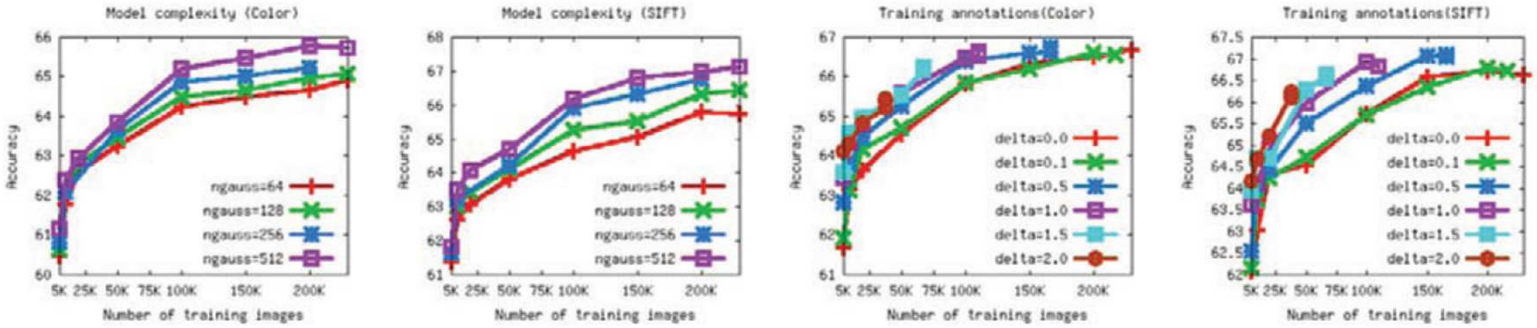

Figure 8. Results for large-scale aesthetic quality categorization for increasing model complexity ((a) and (b)) and increasing values of δ ((c) and (d)).

Content-Based Aesthetic Categorization

- Select images with eight most popular semantic tags – 14368 images.
  - 1. Eight independent SVMs
  - 2. A single, generic classifier with an equivalent number of images.
  - 3. A generic classifier using a large-scale training set composed of 150,000 images.
- The generic large-scale model outperforms the content-based models for all categories using color features, and for 5 out of 8 categories using SIFT features.

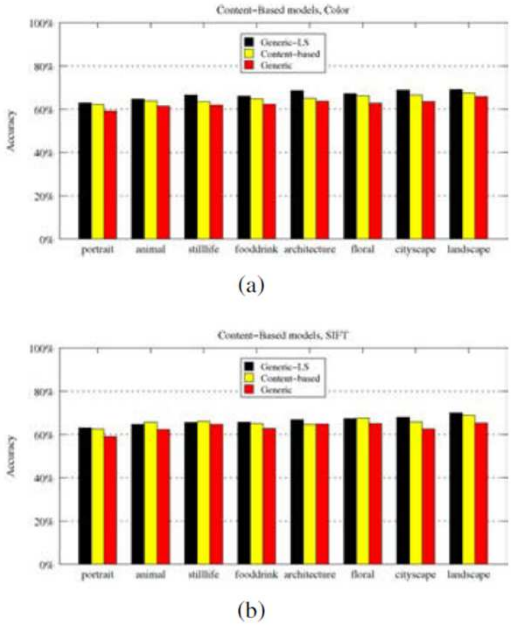

Figure 9. Results of content-based aesthetic quality categorization. Generic models trained on large-scale data out-perform small-scale content-based models.

Style Categorization

- We trained 14 one-vs-all linear SVMs using the 14 photographic style annotations of AVA and their associated images (14,079 images).
- Color histogram feature is the best performer for the “duotones”, “complementary colors”, “light on white”, and “negative images” challenges.
- SIFT and LBP perform better for the “shallow depth of field” and “vanishing point”
- Late fusion significantly increases the mean average precision.

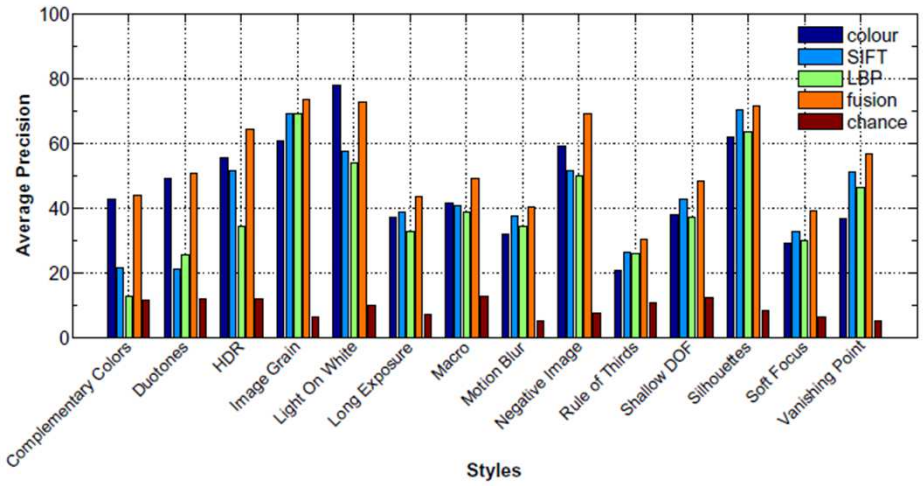

Figure 11. Mean average precision (mAP) for challenges. Late fusion results in a mAP of 53.85%.

Style Categorization

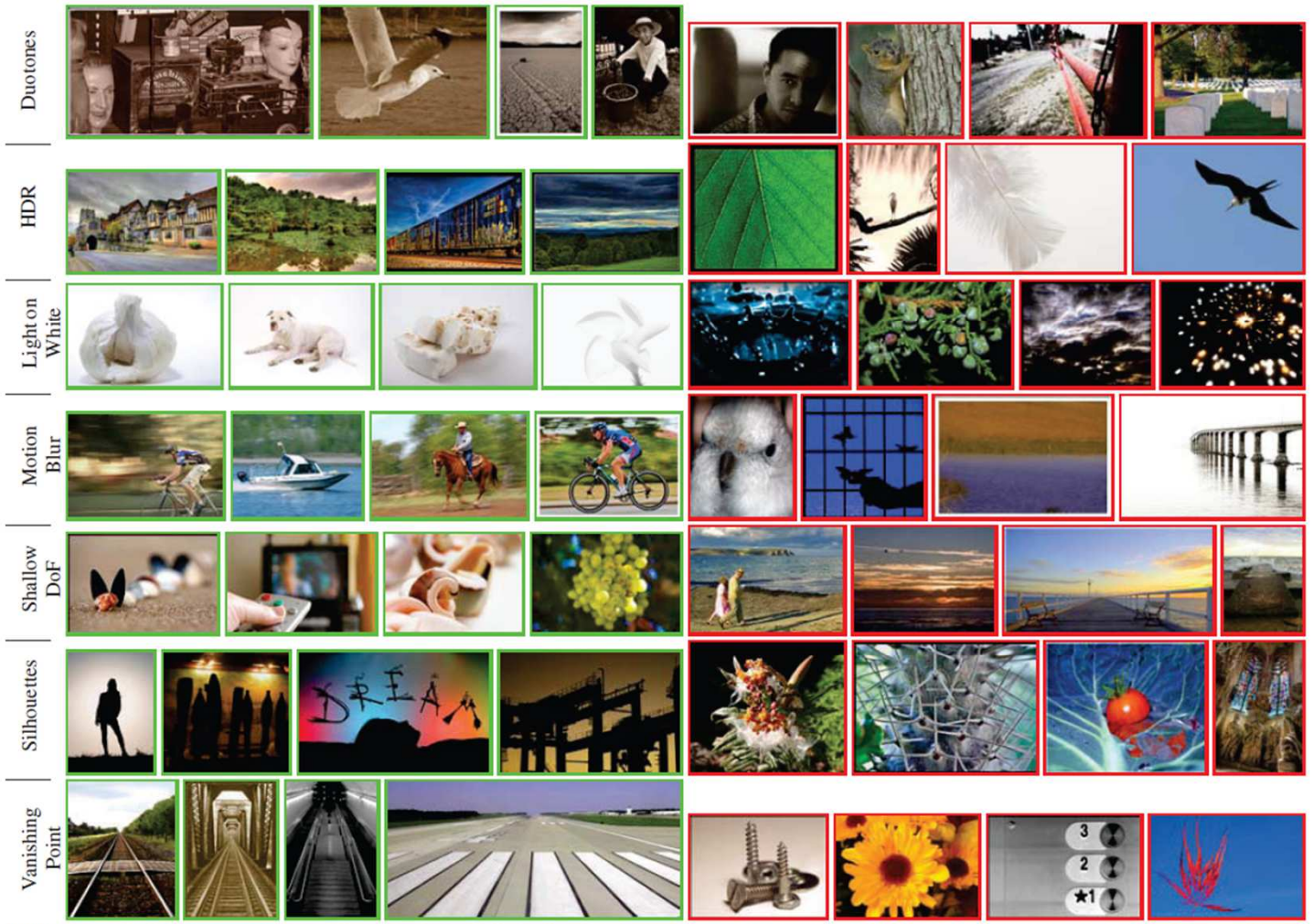

Figure 10. Qualitative results for style categorization. Each row shows the top 4 (green) and bottom 4 (red) ranked images for a category. Images with very different semantic content are correctly labeled.

Discussion and Future Work

- Provide a large-scale benchmark.
- A deeper insight into aesthetic preference.
- Show how richer datasets could help to improve existing applications and enable new ones.