画像修復における定性的・定量的評価法に関する考察
A Study on Qualitative and Quantitative Evaluation for Image Inpainting

河合 紀彦 † 佐藤 智和 † 横矢 直和 †
Norihiko KAWAI Tomokazu SATO Naokazu YOKOYA

1. はじめに

写真についた傷や意図せず写りこんでしまった物体などの画像内の不要部分を取り除き，取り除かれた領域 (以下，欠損領域) を自動的に違和感なく修復することで画像の利用価値を高める画像修復に関する研究が近年盛んに行われている．これまで画像修復に関する多くの手法が発表されてきたが，手法の評価方法については文献ごとに異なり，各々の評価手法の妥当性についてはあまり議論されて来なかった．各文献における画像修復結果の評価法は，以下に述べる 2 種類に大別できる．一方は，従来手法と提案手法による結果画像を主観的に比較することで，提案手法の有効性を示す定性的評価法 [1, 2, 3, 4] であり，他方は，MSE (Mean Squared Error)，RMSE (Root Mean Squared Error)，PSNR (Peak Signal-to-Noise Ratio) などの原画像と結果画像の画素値の差分に基づく定量的評価法 [5, 6, 7, 8] である．画像修復の研究では，個人で利用する写真の利用価値の向上などのように，違和感のない修復を目的とする研究が多く，この場合には必ずしも原画を忠実に再現する必要はない．このような目的に対して，前者のように比較により提案手法の有効性を示すことは妥当であるが，特定の個人の主観のみによる評価であるため，提案手法の有効性を示す根拠に乏しい．それに対して，後者は人の主観が入らず全ての結果に対して同条件で評価を行うことができるが，不要物体を消去するという目的で修復した画像では物体を再現することは無意味であり，原画像と結果画像の差分を用いて結果を評価することはできない．また，上に挙げた例のように，必ずしも原画像を再現する必要がない場合には，定性的評価と定量的評価の結果が一致するとは限らない．

そこで，本稿では，自然な画像修復を目的とした場合において，従来から用いられてきた定性的・定量的評価法がどの程度有効であるのかを明らかにする．具体的には，まず筆者らが提案した画像修復手法 [4] と従来提案されている代表的な 2 つの画像修復手法 [2, 3] を用いて，100 枚の画像 (200 × 200 画素) に対する画像修復実験を行い，それらの修復画像に対して被験者によるアンケート評価を行うことで各手法を定性的に評価する．次に，これらの結果から，アンケート評価を行う際の被験者の人数と実験に用いる画像の枚数について議論する．最後に，RMSE による定量的評価を行い，アンケート評価結果との関係を明らかにする．

2. 従来の評価方法

本節では，従来から画像修復の評価に用いられてきた定性的評価法と定量的評価法について述べる．

2.1 主観的比較による定性的評価

多くの文献では，自然な画像の修復を目的として，生成画像を主観的に比較する定性的評価法が用いられている．このような比較ではエッジのつながりやテクスチャの再現等に着目し，それらの優位点を挙げることで提案手法の有効性を示している．しかし，画像の自然さを判断する基準には人によるばらつきがあるため，特定の個人の主観のみでは，有効性を示すには不十分な可能性がある．また，多くの文献では提案手法の得意とする数枚の画像に対する修復結果を掲載するのみで，従来手法に対する劣位点が述べられていることが少ない．

2.2 原画像と結果画像の差分に基づく定量的評価

いくつかの研究では，原画像を真値とし，結果画像と原画像の画素値の差分に基づく定量的評価が行われている．ただし，このような評価を行うためには故意に原画像に傷やテロップ等による欠損領域を作り，それを修復することで結果画像を作り出す必要がある．定量的評価としてこれまで MSE，RMSE，PSNR が用いられてきた [5, 6, 7, 8]．MSE は以下の式で定義される．

$$MSE = \frac{\Sigma^{N_\Omega}(I_{res}(\mathbf{x}) - I_{org}(\mathbf{x}))^2}{N_\Omega} \tag{1}$$

ただし，I_{res} は結果画像の画素値，I_{org} は原画像の画素値であり，N_Ω は欠損画素数である．また，MSE を用いて RMSE，PSNR はそれぞれ以下のように定義される．

$$RMSE = \sqrt{MSE} \tag{2}$$

$$PSNR = 10\log_{10}(\frac{MAX^2}{MSE}) \tag{3}$$

ただし，MAX は画素値の最大値である．これらの定量的評価では，MSE，RMSE の場合は低い値，PSNR の場合は高い値の方が真値に近い画像であると言える．しかし，画像修復においては原画像を再現しなくても違和感のない画像を作り出すことは可能であるため，自然な画像の修復という目的においては必ずしも真値を求める必要はない．また，文献 [7] では RMSE が必ずしも妥当な評価でないことが示唆されている．以下の実験では，これらの定量的評価値が悪い場合でも，画像修復としては良好な結果が得られる場合があることを示す．

3. 定性的・定量的評価実験と考察

本実験では，様々な特徴を持つ 100 枚の画像 (200 × 200 画素) に対する画像修復を行い，その結果を用いて画像修復手法の評価を行う．以下では，まず実験で用いた手法について概説し，次に，アンケート評価に基づく定性的評価実験について述べる．最後に，RMSE を用いた定量的評価実験について述べ，アンケート評価実験との関係を考察する．

3.1 実験に用いた手法の概説

本節では，評価実験に用いた画像修復の手法について述べる．画像修復手法は，輝度値の連続性を考慮する手

†奈良先端科学技術大学院大学 情報科学研究科
Nara Institute of Science and Technology (NAIST)

法と欠損領域以外の領域のテクスチャを用いる手法に大別できるが，前者は大きな欠損領域を修復した場合，細かいテクスチャが再現できず結果が不鮮明になるという問題があるため，現在，後者が画像修復手法の主流となっている．本実験では，テクスチャを用いる後者の手法の中でも以下に挙げる代表的な手法を実装し，評価実験を行った．本実験で用いた手法は，Criminisi らのテクスチャの逐次合成による手法 [2]，Wexler らのテクスチャの全体最適化による手法 [3]，筆者らが提案したテクスチャの明度変化と局所性を考慮した全体最適化手法 [4] である．以下，それぞれに手法について概説する．

3.1.1 テクスチャの逐次合成による手法 [2]

Criminisi らが提案したテクスチャの逐次合成による手法は，図 1 に示すように欠損領域 Ω 以外の画像内の領域 Φ（以下，データ領域）から欠損領域の境界位置 $\mathbf{x}$ 周辺のテクスチャパターンと最も類似するテクスチャの位置 $\hat{\mathbf{x}}$ を探索し，そのテクスチャを逐次 $\mathbf{x}$ の位置に合成するアプローチを採る．この手法では，テクスチャの合成順に関して，エッジの強い箇所を優先的に合成する方が，単純に欠損領域の周りから逐次合成するよりも不連続なテクスチャが生じにくいという考えから，式 (4) に示すように一定ウインドウ内にある決定済みの画素数 $C(\mathbf{x})$ とエッジの強さ $D(\mathbf{x})$ により計算された $P(\mathbf{x})$ が最も大きくなる位置 $\mathbf{x}$ にテクスチャを合成する．これを，欠損領域が無くなるまで繰り返すことで画像を修復する．

$$P(\mathbf{x}) = C(\mathbf{x})D(\mathbf{x}) \tag{4}$$

この手法では，比較的高速にテクスチャを合成できるが，欠損領域の周りで複雑なテクスチャが多い場合には不連続なテクスチャが生じやすいという問題がある．

3.1.2 テクスチャの全体最適化による手法 [3]

Wexler らは，欠損領域内の画像の尤もらしさを表す関数を定義し，それを最適化することにより画像修復を行う手法を提案している．この手法では，修復結果がテクスチャの合成順に依存せず，欠損領域全体に対して最適な画像を生成することができる．具体的には，欠損領域 Ω とデータ領域 Φ のパターン類似度を用いた式 (5) に示すエネルギー関数 E_w を最小化することで画像修復を行う．

$$E_w = \sum_{\mathbf{x}\in\Omega'} w_{\mathbf{x}}[\sum_{\mathbf{p}\in W} I(\mathbf{x}+\mathbf{p}) - I(\hat{\mathbf{x}}+\mathbf{p})^2] \tag{5}$$

ただし，$w_{\mathbf{x}}$ は画素の信頼度に関する重みで欠損領域の境界に近いほど大きい値になる．また，W は一定サイズのウインドウ，$I(\mathbf{x})$ は位置 $\mathbf{x}$ の画素値である [3]．この手法では，不連続なテクスチャは生じにくいが，テクスチャの明度変化や幾何学的構造変化に弱く，明度の不連続やぼけが発生する場合がある．

3.1.3 テクスチャの明度変化と局所性を考慮した全体最適化による手法 [4]

筆者らは，より多くの画像に対して自然な修復結果を得るために，Wexler らの手法 [3] を基礎として，テクスチャの明度変化と局所性を考慮した全体最適化による画像修復手法を提案している．具体的には，照明変化等に

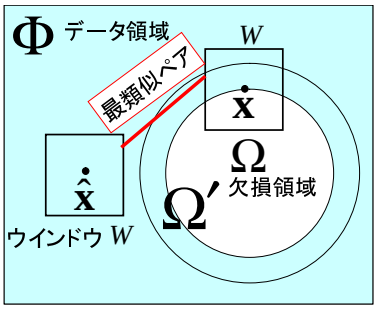


図 1: 画像上の各領域と類似パターンのペアの例

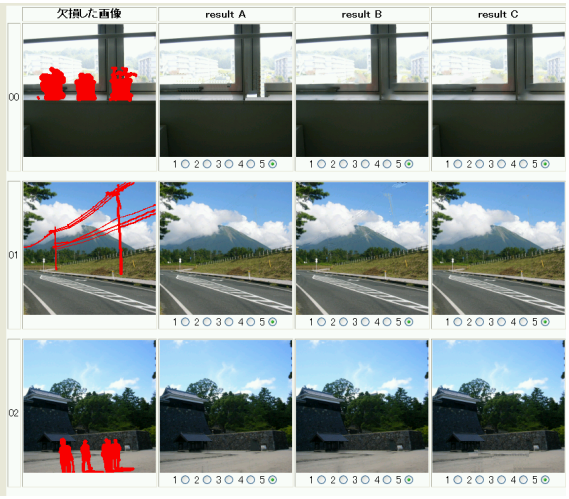


図 2: 修復結果の採点ページ

より同一構造を持つテクスチャでも大きな明度変化があることや，類似したテクスチャ同士が近傍に存在する確率が高いという性質 (テクスチャの局所性) に着目し，式 (6) に示すテクスチャの明度変化を考慮したパターン類似度と局所性を考慮したコスト項 SD によるエネルギー関数 E_k を最小化することで画像修復を行う．

$$E_k = \sum_{\mathbf{x}\in\Omega'} w_{\mathbf{x}}[\sum_{\mathbf{p}\in W} I(\mathbf{x}+\mathbf{p}) - \alpha I(\hat{\mathbf{x}}+\mathbf{p})^2 + w_{dis}SD(\mathbf{x},\hat{\mathbf{x}})] \tag{6}$$

ただし，α は明度変化を許容する輝度補正係数であり，w_{dis} は局所性を考慮したコスト関数の重みである．この手法では，欠損領域全体に対して最適な画像を生成でき，不自然な明度変化やぼけが発生しにくいという特長を持つ．

3.2 アンケート評価に基づく定性的評価実験

定性的評価実験として，上に述べた各手法を用いて得られた修復画像に対して 37 名の被験者による評価を行った．なお，被験者は 20 代前半の男女であり，全員コンピュータを日常的に扱っている．

3.2.1 評価方法

被験者には，アンケート評価のためのウェブページにアクセスしてもらい，図 2 に示すようなレイアウトで並べられた 100 枚の入力画像に対する 500 枚の修復画像に

対して 5 段階の点数評価を行ってもらった．本実験で画像修復に用いた手法は，3.1 節で述べた，テクスチャの逐次的合成による手法 [2](手法 A)，テクスチャの全体最適化による手法 [3](手法 B)，テクスチャの明度変化と局所性を考慮した全体最適化による手法 [4](手法 E) に加えて，手法 E において明度変化のみ考慮したエネルギー関数を用いた場合 (手法 C)，テクスチャの局所性のみ考慮したエネルギー関数を用いた場合 (手法 D) の 5 つとした．修復結果の採点ページ上では，これら 5 つの手法により修復した結果画像を入力画像ごとにランダムな順序で並べ，評価者には各画像がどの手法による修復結果であるのかを知らせなかった．また，本実験では，修復画像を個人のホームページや書籍・雑誌等の写真として利用することを前提として，全く使えない画像を 1 点，十分使える画像を 5 点という回答基準で採点してもらった．

3.2.2 評価結果と考察

手法 A〜E により出力された各々100 枚の修復結果につけられた点数の平均値と各手法が最高点を取った回数を表 1 に示す‡．表 1 から，平均点で比較すると手法 E，手法 D，手法 C，手法 B，手法 A の順で高く，最高点を得た回数で比較すると，手法 E，手法 C，手法 D，手法 A・B の順となった．また，各画像への評価点を用いて有意水準を 5%と設定した t 検定により各手法間を比較することで統計的な裏づけを行った．その結果を表 2 に示す．結果から，手法 C と手法 D には有意な差が見られず，他は有意な差が見られた．従って，画像修復手法の評価として，手法 E，手法 C・D，手法 B，手法 A の順に多くの画像に対して良好な結果を得られる手法だと言え，最高点を得た回数もおおよそそれに一致している．

3.2.3 アンケート評価における被験者数と画像枚数の考察

従来行われてきた少数の画像または少人数の被験者による評価では，画像や人による評価のばらつきにより，手法の優劣が正しく評価されていない可能性がある．そこで，本節では，前述の入力画像 100 枚，被験者数 37 名によるアンケート評価結果における平均点を基準とした手法の優劣を正解とし，実験に用いる画像枚数および被験者数を減少させた場合の結果の信頼性を検証する．

表 1: 100 枚の画像に対する点数の平均点と最高点を取った回数

手法	平均点	回数
手法 A[2]	2.21	7
手法 B[3]	3.24	7
手法 C	3.39	21
手法 D	3.42	20
手法 E[4]	3.60	45

表 2: 有意水準 5%と設定した時の t 検定による各手法間を比較した場合の有意差

	手法 B[3]	手法 C	手法 D	手法 E[4]
手法 A[2]	あり	あり	あり	あり
手法 B[3]	-	あり	あり	あり
手法 C	-	-	なし	あり
手法 D	-	-	-	あり

‡実験に用いた 100 枚の画像と各手法による修復結果とその評価結果を http://yokoya.naist.jp/research2/inpainting/に示す．

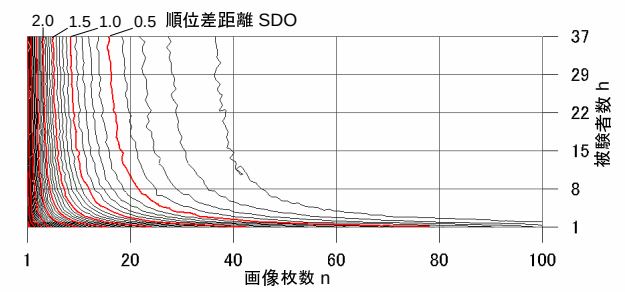


図 3: 画像枚数と被験者数と順位差距離の関係

結果の信頼度を示す指標として，順位差距離 SDO を以下のように定義する．

$$SDO(n,h) = \sum_i \mid f(i,100,37) - f(i,n,h) \mid \quad (7)$$

ここで，$f(i,n,h)$ は画像枚数 n および被験者数 h でアンケート評価を行った際の手法 i $(i \in \{A,B,C,D,E\})$ の順位を表す．本考察では，画像枚数 100 枚，被験者数 37 名の中からランダムに n 枚の画像と h 名の被験者を選び，これを 10000 回施行した平均点による各手法の順位を $f(i,n,h)$ とする．なお，3.2.2 節で得られた結果から $f(i,100,37)$ の値は i が手法 A の時に 5，手法 B の時に 4，手法 C・D の時に 3 または 2，手法 E の時に 1 となる．また，手法 C と手法 D は有意な差がないため，それぞれ 3 か 2 を当てはめて SDO の小さい方を用いる．

図 3 に画像枚数，被験者数，順位差距離 SDO の関係を示す．SDO が 0.1 変化するごとに等値線を引いており，画像枚数 100 枚・被験者数 37 名の時に最も低く $SDO(100,37) = 0$，画像枚数 1 枚・被験者数 1 名の時に最も高く $SDO(1,1) = 5.83$ となった．順位が隣接する手法の順が 1 組だけ入れ替わった場合，SDO の値が 2 になることから，SDO が 0.1 の場合おおよそ 95%の確率で優劣の順位が正解と一致する．図から，被験者数が 37 名の場合，画像枚数が 38 枚以上あれば 95%以上の確率で正解と一致し，画像枚数が 100 枚の時，被験者が 3 名いれば 95%以上の確率で正解と一致する．しかし，例えば $n = 3$，$h = 3$ の場合には，$SDO = 3.34$ となり SDO の値が 2 を上回ってるため、手法間の順位が平均的に 1 組以上入れ替わっていることがわかる．以上のことから，数人による数枚の画像に対する主観評価では，多くの被験者および多くの画像を用いた場合の評価と結果が異なる可能性が高い．

3.3 定性的評価と定量的評価の関係

画像修復の定量的評価として従来，修復画像と原画像の画素値の差分に基づいて計算される MSE，RMSE，PSNR が用いられてきたが，本実験では RMSE を用いる．なお，本実験で用いた 100 枚の画像のうち 71 枚の画像においては，特定の物体を取り除くために欠損領域の指定を行っ

ているため，その物体を復元することは無意味である．したがって，物体の位置に関わらず欠損領域の指定を行った残りの 29 枚に対して RMSE の評価を行った．

表 3 に各手法による 29 枚の結果画像に対する RMSE の平均値とアンケート評価の平均値を示す．また 29 枚の画像に対する RMSE とアンケート評価の間の相関係数の分布に関して 5 手法を用いた場合を図 4(a) に，手法 A を除いた 4 手法を用いた場合を図 4(b) に示す．なお，本実験で用いた相関係数 Cc は以下の式で計算される．

$$Cc = \frac{N_m \Sigma_i^{N_m} R_i S_i - \Sigma_i^{N_m} R_i \Sigma_i^{N_m} S_i}{\sqrt{N_m \Sigma_i^{N_m} {R_i}^2 - (\Sigma_i^{N_m} R_i)^2}\sqrt{N_m \Sigma_i^{N_m} {S_i}^2 - (\Sigma_i^{N_m} S_i)^2}} \tag{8}$$

ただし，N_m は比較に用いた手法の数であり，R_i，S_i はそれぞれ RMSE の値，被験者による評価値の平均値である．

表 3 と図 4(a) から RMSE とアンケート評価の間には強い負の相関関係が確認できる．しかし，図 5 に示す 3 枚の画像においては，正の相関関係が見られた．これら 3 枚の画像は他の画像より比較的周波数が高いテクスチャを持つ画像である．文献 [7] でも指摘されているが，現在定量的評価に用いられている MSE，RMSE，PSNR は 2 つの画像間の対応する画素値の差分を用いて計算されるため，生成される画像の見た目の違和感に関係なく，高周波テクスチャの位相のずれに対して敏感に値が変化する傾向がある．従って，RMSE は高周波を多く含む画像の評価には適していないと言える．図 4(b) は RMSE とアンケート評価の評価値が顕著に低い手法 A を除いた 4 手法による相関係数の分布であるが，約 3 分の 2 の画像については負の相関関係，約 3 分の 1 の画像については正の相関関係を示している．これは，RMSE の微小な差が必ずしも画質の差を表していないことを示している．以上のことから，RMSE は人間の主観による画質の評価と一致することも多いが，画像修復結果の絶対的な判定指標として用いることは難しい．

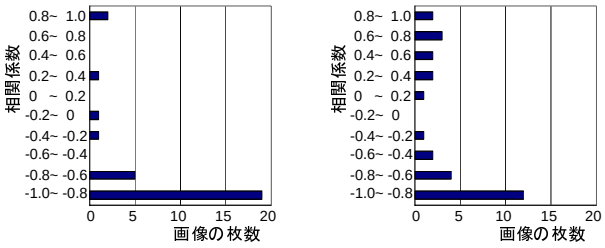
4. まとめ

本稿では，100 枚の入力画像を用いた被験者 37 名によるアンケート評価に基づく定性的評価と RMSE に基づく定量的評価を行うことで画像修復手法の評価法に関する考察を行った．また，アンケート評価と RMSE の関係について議論した．多数の入力画像や被験者を用いたアンケート評価を行うことで，個人の主観や画像の特徴によるばらつきに左右されない画像修復手法の評価を行うことができることを示した．また，RMSE による定量的評価は，画像修復結果の評価に必ずしも適さないことを示した．

今後は，定性的評価と一致する定量的評価法を検討する必要がある．また，評価に用いる画像を画像の特徴により分類することで，画像修復手法をより詳細に評価する手法を検討する．

表 3: 29 枚の画像に対する RMSE とアンケート評価それぞれの平均値

	RMSE	アンケート評価値
手法 A[2]	42.95	1.99
手法 B[3]	28.40	3.11
手法 C	27.83	3.44
手法 D	28.36	3.34
手法 E[4]	27.44	3.66



(a) 手法 A〜E による結果　　(b) 手法 B〜E による結果

図 4: RMSE とアンケート評価の間の相関係数と画像の枚数の関係

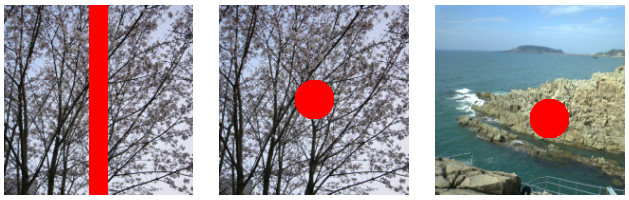


図 5: RMSE とアンケート評価の間に正相関が得られた画像

参考文献

[1] M. Bertalmio, G. Sapiro, V. Caselles and C. Ballester: “Image Inpainting,” Proc. ACM SIGGRAPH2000, pp. 417–424, 2000.

[2] A. Criminisi, P. Pérez and K. Toyama: “Region Filling and Object Removal by Exemplar-Based Image Inpainting,” IEEE Trans. on Image Processing, Vol. 13, No. 9, pp. 1200–1212, 2004.

[3] Y. Wexler, E. Shechtman and M. Irani: “Space-Time Completion of Video,” Trans. on Trans. on Pattern Analysis and Machine Intelligence, 29, No. 3, pp. 463–476, 2007.

[4] 河合，佐藤，横矢: “パターン類似度に基づくエネルギー最小化による画像修復”, 電子情報通信学会技術研究報告, PRMU2006-163, pp. 13–18, 2006.

[5] 長, 本田: “局所的と大局的処理の二段階画像インペインティング”, 電子情報通信学会 技術研究報告, PRMU2006-279, pp. 143–148, 2007.

[6] 天野, 佐藤: “固有空間法を用いた BPLP による画像補間”, 電子情報通信学会誌 D-II, Vol. J85-D-II, No. 3, pp. 457–465, 2002.

[7] 天野, 佐藤: “kBPLP 法を用いた高次元非線形射影による画像補間”, 電子情報通信学会誌 D-II, Vol. J86-D-II, No. 4, pp. 525–534, 2003.

[8] T.K. Shih, R.C. Chang, L.C. Lu and L.H. Lin: “Large Block Inpainting by Color Continuation Analysis,” Proc. Int. Multimedia Modelling Conference, pp. 196 – 202, 2004.