巻頭言

評価用データセット

斎藤英雄
慶應義塾大学

我々が書く技術論文は，大抵の場合，新規な何かを提案したものである．提案しただけでは技術論文にはなりにくく，通常は，次の二つを実施することが必要である．

・比較：関連する研究として既発表・既存の関連研究を挙げ，それらと「比較」することにより，新規提案の優位性を示すこと．
・評価：その新規提案の性能や有効性を確認するために実験等を実施し，その結果を「評価」すること．特に，この「評価」は，客観的な比較のためには必須となる．

「比較」については，行われないことも稀にある．これはその論文で提案していることが斬新で，同じ目的を目指した提案が存在しないことが明白であったり，同様な結果が誰にも得られていなかったりすることが明確な場合に限られる．たいていの場合は，同じ目的の提案がすでにいろいろと提案されており，同様な結果は他手法でも得られていることが多い．このような場合は，提案手法による結果を示しただけでは，決して論文はアクセプトされない．そこで，何らかの関連手法と比較して，優位性や有効性を示す必要が出てくるのが通常である．この比較の際，新規提案によって得られた結果を示し，さらに，関連手法により得られる結果を示すか引用するだけで，提案手法の優位性を示すことが可能なこともある．しかしながら，大抵は，そのような比較だけでは「比較に客観的な基準がないので，本当にそっちが良いのかわからない」とか言われてしまうので，結果を「評価」する仕組みが必要になってくる．

この「評価」の方法は，論文が扱っているトピックに応じていろいろなやり方で行われている．その典型的な方法としては下記のようなものが挙げられる．

（1）正解が既知な状態での評価

本来正しい数値，信号，画像（Ground　Truth）があって，それに対して，論文で示された手法により得られた結果との差分（つまり，端的に言えば誤差）の大きさにより結果を評価する．主観の入らない客観性の高い評価が可能である．

（2）正解がないが，実験を通して比較評価

正解を考えることが無意味で，人間主観や感性により評価するしかない場合に，評価者に使わせたり見せたりして，そのときのアンケートで性能を数値化する方法．評価者の違いによる揺らぎを防ぐために複数の評価者を準備するのが普通であり，大学の研究では，20 代前半の学生 10 ～ 20 名程度に対してアンケートをすることが多い．

（1）ができれば理想的である．これが容易な分野としては，例えば，画像符号化の分野がある．符号化したい正解画像が与えられ，符号化と復号化の結果得られる画像と，正解画像を比較すれば，簡単にその符号化手法の性能を評価できる．しかし，通常は画像が変われば性能が変わる．そこで，公平な評価をするための性能を評価する指標となる「標準の画像」が作られた．画像符号化の分野では，有名な LENA の画像がある．LENA は当時のグラビア写真をスキャンした画像であり，被写体のモデルさんの肩から上の画像は有名だが，オリジナルはその下もあることは一部の研究者にはよく知られている事実である．（興味がある人は，LENA やレナで検索してみると良い．面白い話（いわゆるトリビア）がたくさん詰まっている．）

このようなデータは，画像検索・認識向けにも標準的なものが使われている．例えば，CALTEC101 という画像データベースには，101 種類のカテゴリ群の画像がカテゴリー毎に 40 から 800 枚含まれているものであり，「一般物体認識・画像カテゴリ認識」といった研究の学習データ・性能評価データとして非常に広く利用され，この分野の技術の進展に大きく貢献した．

2 台の視点の異なるカメラで得られた画像から距離画像を得るための「ステレオ法」は古くからコンピュータビジョンの典型的な問題として膨大な研究がなされ

てきたが，それを同じデータセットで評価するために Middlebury Stereo Datasets というのが広く利用されている．これは，正解の距離画像と，そのシーンを撮影したステレオ画像のセットであり，これを使って，正解の距離画像と，自分で考案したステレオのプログラムで生成した距離画像を比べた評価値から，それがほかの方法と比べてどの程度良いのか，ということが比較できる．

上記の LENA(に代表される標準画像データ群), CALTEC101，Middlebury Stereo Datasets といった標準データを使うことのメリットは，単に評価が容易・可能というだけではない．このデータベースで自分の考案した認識プログラムの性能を評価するだけで，自動的に他の関連手法との比較・競争ができることである．他の関連研究においても同じデータベースを使って，同じ評価をしているため，共通のデータで評価を行っている限り，比較対象の方法を自分で実装し実験する必要がなくなり，比較対象の論文から認識率等の数値を拾ってきて，比較しさえすれば，立派に客観的な比較評価を行うことができ，それを論文に載せれば，客観的評価による比較により有効性が示されている，として，論文が採択される可能性は増大するように思える．

このような評価用データの存在により，多くの研究が客観的に同じ土俵の上で比較できる環境が整うと，そこに参入した研究者は比較が容易になるので，論文が書きやすくなるし，さらにたとえ何回も新たな研究に負けようとも，必ず引用されるので被引用回数で価値が評価される論文の価値がどんどん上がる．それを見た周辺の研究者も参入し，・・・と好循環でその分野が発展するのである．

さて，このようなデータセットとして，私が現在委員長を務めている「複合現実感研究会」では，数年前から TrakMark と名付けたワーキンググループにおいて，複合現実感研究のための評価用データセットの整備を進めている．この WG は，この研究会の創設者である立命館大学の田村秀行先生が中心になって，先に述べたような思い，つまり，評価用のデータセットにより，当該分野が大きく進展する，ということを期待して数年前からスタートした．

複合現実感の研究分野では，例えば AR Toolkit Marker により実現可能な動画像からのカメラの位置姿勢のリアルタイム推定を，マーカレスで，しかも被写体がマーカのような平面物体でなくても実現できる手法が盛んに研究されており，現在でも重要な研究要素技術となっている．しかし，この複合現実感のためのカメラ位置姿勢のリアルタイム推定問題に対しては，大抵の場合，研究者が独自に撮影した動画像を使うのみであり，関連研究と比較しようとしても，それを自分で実装するか，そのプログラムを提供してもらわなければならず，客観的な比較が難しかった．このため，結局のところ，結果として得られる複合現実感提示映像の見た目による評価に頼る，といったことも少なくなかった．

そこで，TrakMark WG では，このためのいろいろな研究を客観的指標により評価・比較可能にすることを目指したデータセットを作成し公開している．いくつかの典型的シーンに対する動画像シーケンスと，それにより推定したカメラの位置姿勢の精度を客観的に評価するためのカメラの位置姿勢の正解データも提供している．さらに今後は，比較評価を容易にするために，評価の指標やその評価指標に基づく評価値を算出するプログラムも含めて提供したいと考えている．この評価用データセットについては，これまで，複合現実・拡張現実に関する国際会議 (ISMAR) 等でワークショップを開催するなど，宣伝に努めてきたものの，まだ多くの研究者に使われる状態にはなっていないのが実情である．そこで，今年は，9 月 29 日から 10 月 3 日に福岡で開催される国際会議 ISMAR2015 においてコンテストを企画したり，国内の画像認識分野で 20 年の歴史を誇る電子情報通信学会の PRMU 研究会の行っているコンテストと連携する等して，データセットのユーザを増やしたいと思っている．

後半は，TrakMark WG で整備中のデータセットを「主観的」評価により紹介してしまったので，このデータセットの意義を「客観的」に評価するデータセットがさらに必要では？という疑問を持たれないうちに，この巻頭言を終わりたい．

【略歴】
斎藤英雄 (SAITO Hideo)
慶應義塾大学　教授（理工学部　情報工学科）
1992 年，慶應義塾大学大学院理工学研究科電気工学専攻博士課程修了．同年，同大学助手．その後専任講師，助教授を経て，2006 年，同大学理工学部情報工学科教授． この間，1997 年～ 1999 年まで，学術振興会海外特別研究員として，米国カーネギーメロン大学ロボティクス研究所に滞在し，主に Virtualized Reality の研究に従事．2000 年～ 2003 年，JST さきがけ研究「情報と知」領域研究員兼務．2006 年～ 2012 年，JST CREST 研究代表者．現在，主としてコンピュータビジョンとその VR 応用に関する研究等に従事．博士 (工学)．正会員．