Evaluation in Computational Creativity

(Some views.
Anna Jourdanous provides a wider perspective)

Evaluation is important – and difficult

- Evaluation of creativity allows us to compare methods and control progress
- However, evaluation of creativity is very difficult
  - No precise definition of creativity
  - Various goals (novelty, value, originality, ...)
  - Context-dependence
  - Cost of evaluation
  - Evolution of (social) systems
  - ...

What to evaluate?

- Machine creativity:
  Creative performance of creative programs
- Computer-supported creativity:
  Increase in creativity of humans using CC tools
- Creativity studies: Increase in knowledge about creative processes

- Focus here: evaluation of machine creativity

Evaluation of Machine Creativity

Two possible targets in evalution of machine creativity (Colton 2008):

- Artefact-based evaluation: are the results creative?
  - e.g: novelty and value of results
- Process-based evaluation: is the process creative?
  - e.g: combinatorial/ exploratory/ transformational creativity; generation vs. creativity by Ventura; creative acts of the FACE model

Ritchie's Framework for Artefact Based Evaluation

Ritchie (2007)

Essential properties

Consider a set R of artefacts produced by a system.

Primitive properties that can be considered:

– **Typicality**: Is the artefact a typical/ recognizable example of the target genre?

– **Novelty**: How (dis)similar is the artefact to existing examples of its genre?

– **Quality** [= Value]

Formal definitions

- typ(a) = amount of typicality associated to artefact a
- val(a) = amount of quality associated to a
- $T_{\alpha,\beta}(X) = \{a \in X \mid \alpha \leq typ(a) \leq \beta\}$
  - Set of artefacts a with typicality between α and β
- $V_{\alpha,\beta}(X) = \{a \in X \mid \alpha \leq val(a) \leq \beta\}$
  - Set of artefacts a with value between α and β
- size(X) = number of elements of X
- ratio(X,Y) = size(X) / size(Y)

Some criteria

Criterion 2 $\mathit{ratio}(T_{\alpha,1}(R), R) > \theta$

- at least fraction θ of results R have high typicality (>α)

Criterion 4 $\mathit{ratio}(V_{\gamma,1}(R), R) > \theta$

- at least fraction θ of results R have high value (>γ)

Criterion 5 $\mathit{ratio}(V_{\gamma,1}(R) \cap T_{\alpha,1}(R), T_{\alpha,1}(R)) > \theta$

- at least fraction θ of results R have both high value (>γ) and high typicality (>α)

Inspiring set

- Any creative system is based on some existing examples, in one way or another. These can – and should – be taken into account.
- The *inspiring set* consists of all the relevant artefacts known to the program designer, or items which the program is designed to replicate, or a knowledge base of known examples which drives the computation within the program
- Inspiring set ≈ training set in ML/DM

Some more criteria

Criterion 9 $ratio(I \cap R, I) > \theta$

- Results R reproduce at least fraction θ of the inspiring set I
- Is the system able to reproduce its training examples?

Criterion 10 $ratio(R, I \cap R) > \theta$

- Results R contain at least θ-1 times as many items outside the inspiring set I as inside it
- Can the system extrapolate outside the training examples?

Novelty vs. typicality?

Novelty and typicality are subtly different:

- Not recognizable as a member of the genre → low typicality
- Very different from the inspiring set (but possibly very clearly within the genre) → high novelty

Comments

Note: Ritchie does not prescribe a set of criteria. Instead, the criteria must be designed and chosen according to the goals and needs of each work; Richie gives examples of some of the possible criteria that one may want to use.

FACE Model for Process-Based Evaluation

Pease and Colton (2011)

F, A, C, E

- Focus on *creative processes*, not their results
- In the FACE model, systems can be characterized by their creative acts
- The four aspects of the model:
  - F – framing
  - A – aesthetics
  - C – concept
  - E – expression
- Here we present a simplified version

FACE aspects

- C: the *concept* or the idea of the artefact
  - E.g. use of excessive rhyming in poetry
- E: a concrete *expression* of the concept
  - E.g. a poem that uses excessive rhyming
- A: a measure of *aesthetics* of the work of art
  - E.g. emotionality etc. of a poem
- F: background information about the piece (*framing*)
  - E.g. a description of why excessive rhyming could be interesting, and what the poem expresses

Framing

- Framing is especially important for computational creativity
- It is difficult to appreciate the output (expression) without knowing anything about the process, its goals, etc.
  - E.g., is the resulting image pretty just by chance? Or did the system produce it based on some specific criteria and goals? Was the process complicated? Is there some intention, e.g., a message that is being conveyed?

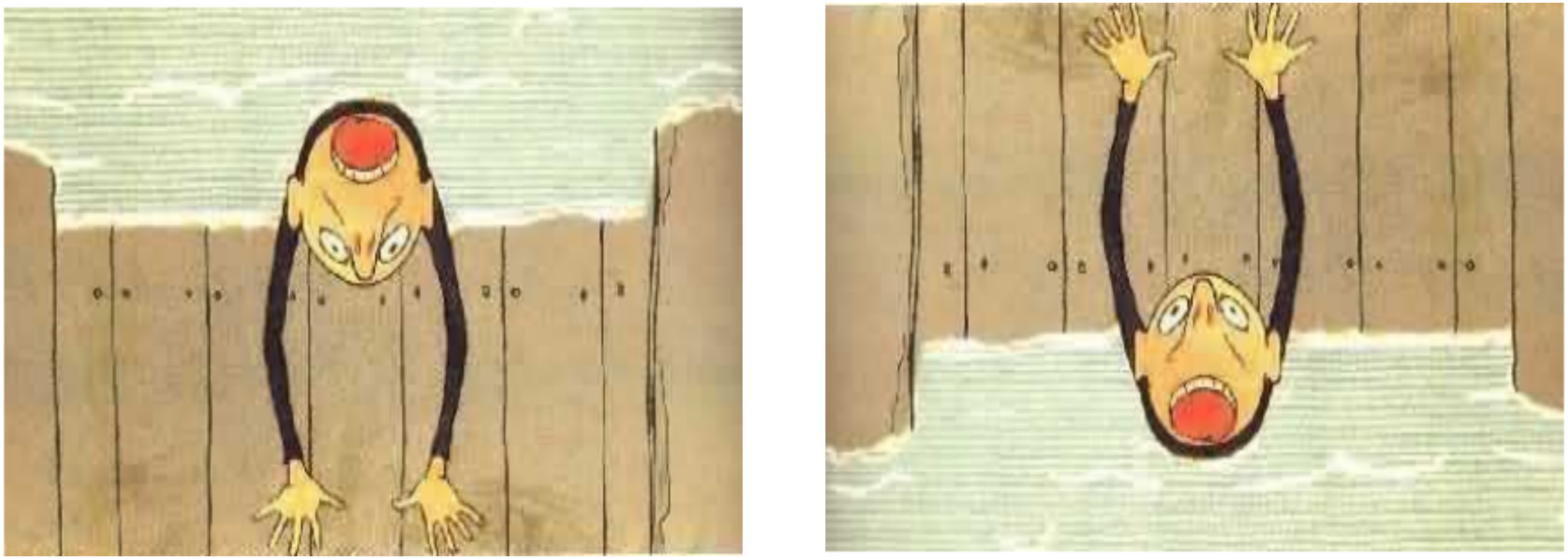
Ground level of FACE

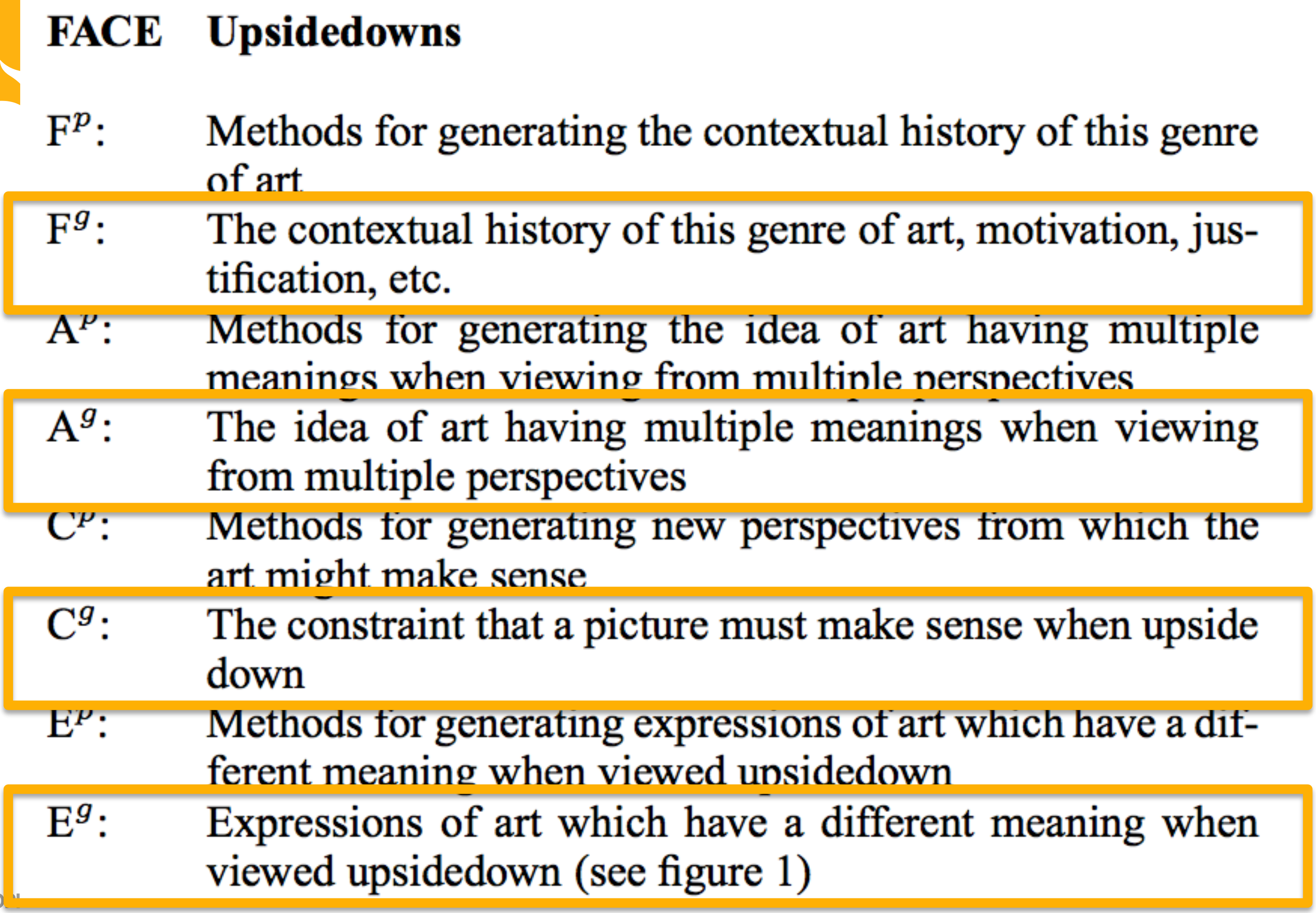
- Ground-level generative acts and their products
  - Act F^g → generates an item of framing information
  - Act A^g → generates an aesthetic measure
  - Act C^g → generates a concept
  - Act E^g → generates an expression of a concept
- Any system can now be described in terms of who carries out these acts, and how
  - A simple generative system only performs E^g
  - A system that learns to evaluate also performs A^g
  - (The programmer and other humans probably perform the other acts)

Meta-level of FACE

- FACE also has a meta-level: processes that produce ground-level generators
- Process-level acts and their outputs:
  - Act F^p → generates a method for generating framing information
  - Act A^p → generates a method for generating aesthetic measures
  - Act C^p → generates a method for generating concepts
  - Act E^p → generates a method for generating expressions of a concept

Example from Pease et al, 2011
The Upside Downs by Verbeek





FACE	**Upsidedowns**
F^p:	Methods for generating the contextual history of this genre of art
F^g:	The contextual history of this genre of art, motivation, justification, etc.
A^p:	Methods for generating the idea of art having multiple meanings when viewing from multiple perspectives
A^g:	The idea of art having multiple meanings when viewing from multiple perspectives
C^p:	Methods for generating new perspectives from which the art might make sense
C^g:	The constraint that a picture must make sense when upside down
E^p:	Methods for generating expressions of art which have a different meaning when viewed upsidedown
E^g:	Expressions of art which have a different meaning when viewed upsidedown (see figure 1)