Automated video games evaluation based on the template formalism

François Nel, Marc Damez, Nicolas Labroche and Marie-Jeanne Lesot
LIP6 -Université Pierre et Marie Curie-Paris6, UMR7606,
104 avenue du président Kennedy, Paris, F-75016 France
francois.nel@lip6.fr

Abstract

This paper proposes a new approach for video game evaluation, based on a hierarchical model of quality criteria, in the framework of the template formalism: the latter offers a principled aggregation methodology to combine elementary and directly evaluable criteria into more complex properties until the global quality of a game can be assessed. We propose a structured organisation of existing video game quality criteria, and its implementation in the template formalism. The first experiments performed with real data show this model constitutes a relevant and powerful tool that provides interpretable assessment results consistent with expert evaluations.

Keywords: template formalism, video games, computer games, evaluation, user testing.

1 Introduction

The intent of this paper is to present an automatic method to evaluate the quality of video games. This work is a response to a need motivated by the increased competition among game developers. In order to subsist, game companies have to develop high quality games, therefore they are looking for efficient methods to evaluate their games.

There already exist studies about the automatic evaluation of video games, based on the adaptation of software quality evaluation [6, 3, 5] or focusing on specific aspects of video games [1, 12, 13]. Yet these works usually consist of lists of quality criteria that must be assessed. These lists tends to be as complete as possible, nevertheless there is usually no indication on how to use these criteria to deduce an evaluation of the game. When specified, the quality of the game is defined as the average quality for all criteria [12] and does not take into account the complexity and specificity of each game type.

In this paper, we propose a new model for the game quality evaluation, based on a hierarchical organization of the criteria, in the template formalism: this formalism, that was initially designed for scenario pattern recognition [2, 9, 10], consists in recursively breaking up complex phenomena into less complex ones, until elementary phenomena that can be directly observed are reached.

In the case of the video game evaluation, the principle is to decompose the notion of a high quality game into specific properties, derived from the existing evaluation heuristics that may be difficult to assess directly, but that can in turn be decomposed into more specific criteria, until directly evaluable criteria are obtained. This breaking up of notions into a hierarchy and the aggregation tools to then combine these concepts make the template formalism very appropriate in this context.

Furthermore, the template formalism offers many advantages for video game evaluation:

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1567–1574
Torremolinos (Málaga), June 22–27, 2008

first, it allows to integrate linguistic variables, which makes it possible for the user to explicitly name positive and negative aspects of the game. Thus it gives an intelligible and interpretable evaluation that can be explained, understood or questioned.

Secondly, our approach to video game evaluation is based on user testing and our goal is to translate as clearly as possible the intuitive reasoning of experts. The template formalism offers fuzzy mechanisms that are particularly relevant to achieve this aim.

Lastly, video games are usually categorised into different types. Some of them, like FPS (First Person Shooter) or RTS (Real Time Strategy) have very specific and distinct characteristics, and should thus be evaluated differently. Therefore the recognition ability of the formalism which was first used as a scenario recognition tool [2] is central in our process. It can be used to distinguish the type of a game in order to adapt the evaluation process.

The paper is organised as follows: in Section 2, we first present the proposed organisation of evaluation criteria that are used in the template formalism. In Section 3, the template model for video game evaluation is described in more details, along with the chosen aggregation operators and parameters to be set. Section 4 describes the prototype that implements these principles, and presents preliminary results obtained for 3 FPS games through the prototype. Lastly, Section 5 concludes and presents perspectives.

2 Criteria for video game evaluation

2.1 State of the art

There exist many studies investigating video game quality. Some focus on the adaptation of software evaluation heuristics for the game development [6, 3, 5]. Other consider how players experience video game through specific points of view such as game immersion [1], the player enjoyment [12] or the role of competition [13].

More generally, game heuristics are usually organised into four categories: the *gameplay* criteria deal with the challenges, problems, obstacles, puzzles the player faces during the game. The *game story* criteria concern the story, plot and character development. The *game mechanics* criteria evaluate the way the player interacts with the game world and environment. Lastly the *game interface* criteria assess the quality of the set of tools the player uses to interact with the game.

The practical exploitation of these heuristic lists is usually not described; when specified, it appears that the global quality of a game is computed as the average of the individual heuristic evaluation [12].

In this paper, we propose a hierarchical model that provides a structured organisation of the criteria and makes it possible to distinguish between elementary properties and complex ones. Moreover, we choose to distinguish two types of criteria, depending on whether they are general and can apply to any video game (generic criteria) or whether they are dependant of the type of the game (type-specific criteria). In the following, we detail each type in turn.

2.2 Generic criteria

The generic criteria are those classically described in the literature and mentioned in the previous subsection. They can be applied to any type of video game and are, as precised in Section 2.1, generally organised into four categories. They correspond to the four main aspects to be taken into account for game evaluation.

We propose to compile the game heuristics into a precise, well-classified organisation based on these categories. The overall compilation groups more than seventy criteria, into seventeen intermediate subcategories and four categories. Figure 1 details the organisation of the *Game mechanics* category, the overall compilation as well as relations between categories or subcategories (e.g. some "Controls" criteria in Figure 1 are related to "Intuitivity" concepts) are not presented here for readabil-

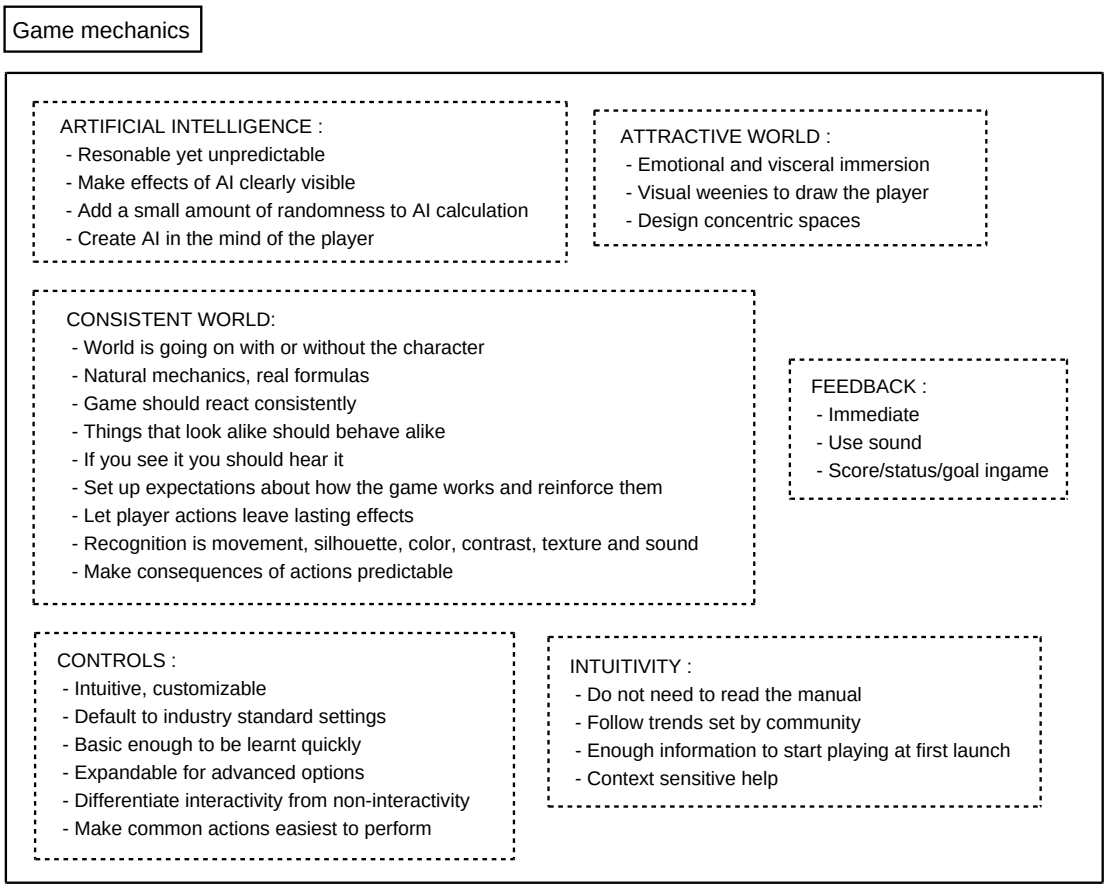

Figure 1: The *Game mechanics* criteria

ity reasons. The following example illustrates the construction process.

Let us consider the following heuristics generally classified in *game mechanics*: "Use real world formulas" [5], "Mechanics should feel natural and have correct weight and momentum" [6], "The world is going on whether your character is there or not" [6], "The player feels as though the world is going on whether their character is there or not" [3]. We realised that they are all about the consistency of the world in the game. So we introduce a subcategory named *Consistent world* and two basic criteria: "World is going on with or without the character" and "Natural mechanics, real formulas". These criteria contribute to define and enrich the introduced subcategory. The template formalism presented in Section 3 includes aggregation tools to exploit this representation.

2.3 Type-specific criteria

In addition to the previous criteria, type-specific criteria are introduced to be able to adapt the evaluation process to the type of game. These are *specific* because they characterize a type of game and have to be met for the type to be recognized. For example, in the case of a FPS, we realised that the reactivity of the player actions during the game has to be very high or that the learning period of the controls has to be short. These two criteria are specific to the FPS type, thus they will play a major part in the evaluation of a FPS game as type-specific criteria.

3 Template-based proposed model

As formalism to model the previous hierarchical structure of the criteria, we propose to use the template formalism described in [2, 9, 10] and introduced for scenario pattern recogni-

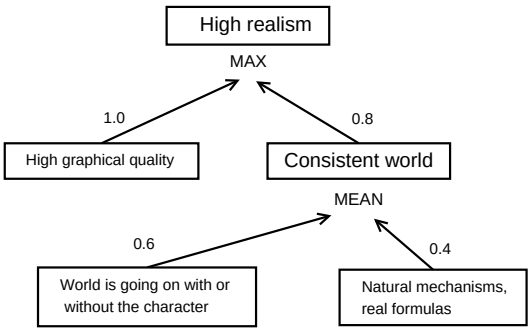

Figure 2: The "High realism" template

tion. The general principle of a template is the breaking up of a complex phenomenon into a combination of less complex phenomena, until elementary phenomena that can be directly observed from the data are reached.

3.1 The template formalism

The template formalism is based on a tree structure. The hierarchical breaking up of concepts is implemented in the tree using the usual splitting of one node (parents) into multiple sub-nodes (children). The process of deduction of a node from its children is implemented by a selection of aggregations operators. In the formalism, weights can be assigned to each sub-phenomenon to express its relative importance during the evaluation process. Figure 2 illustrates an (incomplete) template using the example given in Section 2.2 and presents two forms of aggregation "MAX" and "MEAN".

In the formalism, the leaves of a template are associated with constraints expressed using a fuzzy representation. The evaluation of a leaf is defined as a compatibility degree that represents the extent to which the constraint of the leaf is verified. The propagation of these degrees in the template up to the root infers the global evaluation.

For example, on Figure 2, the constraint of the criterion "Graphical quality" is expressed as "high" and is represented by a fuzzy set defined on a normalised scale between 0 and 10.

3.2 The video game template

In the case of video game evaluation, the root of the template is associated with the concept of "high quality game" of a certain type. Leaves are atomic evaluation criteria that should be understandable by players and the other nodes are intermediate criteria or characteristics.

The goal is to evaluate a game described in terms of elementary properties. This is done by evaluating the extent to which these properties match the complex concept, and then propagating the scores up the tree until its root is reached.

Thus, the final evaluation of the template depends on the evaluation of the two types of criteria (generic and type-specific) and the nodes of the template. Criteria evaluation and node evaluation are described in turn in the following subsections.

3.2.1 Criteria evaluation

The template leaves evaluation depends on their type: leaves representing generic criteria, such as the leaves 'World is going on with or without the character" and "Natural mechanics, real formulas" on Figure 2, are evaluated by a player with a simple mark over 10. It is a very simple way to evaluate a criterion which does not imply any fuzzification process.

In the case of type-specific criteria evaluation, we propose to use a more flexible evaluation: the player is allowed to give either a simple note (as for generic criteria) or a fuzzy evaluation to characterize the criterion. The process suggested in the template formalism [2] to evaluate constraints is the calculation of a compatibility degrees. Indeed, the type-specific criteria, such as the "High graphical quality" leaf on Figure 2 are modeled with fuzzy sets that represent constraints to be met. For such leaves, the evaluation is based on the computation of a degree that indicates the extent to which the observed game characteristic matches the constraint.

In other words, one must choose a compati-

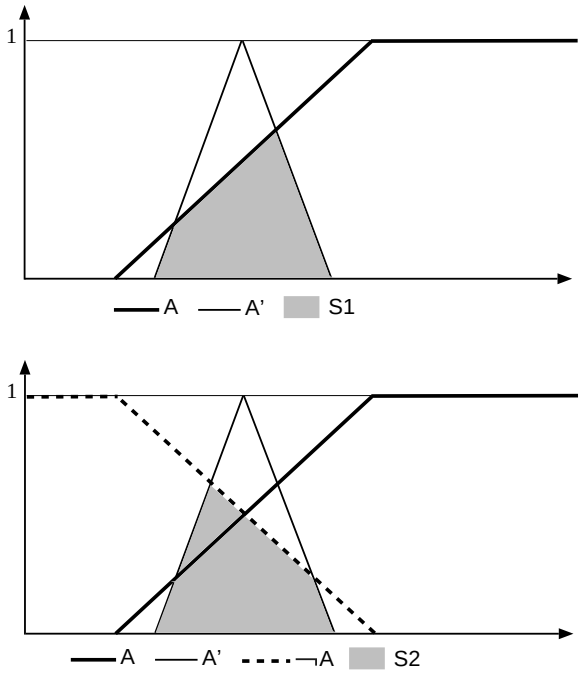

Figure 3: Isomoto compatibility measure

bility measure S that basically computes the satisfaction degree between two fuzzy sets A being the reference set (i.e. the representation of the constraint in the template) and A' the set to compare with A (i.e. the player observation). As exposed in [4], the Isomoto compatibility measure [8] is appropriate given the fact that the measure must intuitively match the following properties:

1. $0 \leq S(A, A') \leq 1$
2. S is increasing with $|A \cap A'|$
3. S is decreasing with $|A' - A|$
4. S does not depend on $|A - A'|$
5. $S(A, A') = 0$ if $A \cap A' = \emptyset$
6. $S(A, A') = 1$ if $A' - A = \emptyset$ and $A \cap A' \neq \emptyset$

Considering the areas $S_1 = |A \cap A'|$ and $S_2 = |\neg A \cap A'|$ as illustrated in Figure 3, the Isomoto compatibility degree is given by:

$$S(A, A') = \frac{S_1}{S_1 + S_2}$$

One can note that this measure is not symmetric. Moreover, if A and A' are non-fuzzy sets, the Isomoto compatibility degree is the satisfiability measure:

$$S(A, A') = \frac{|A \cap A'|}{|A'|}$$

In the case of the "High graphical quality" in Figure 2 for instance, the criterion "graphical quality" has been formulated by experts as "High". Let's suppose that the player defines the graphical quality of the game to evaluate as "Very high". The evaluation of this leaf in the template is then the Isomoto compatibility between "Very high" and "High".

3.2.2 Node evaluation

When all children of one node have been evaluated, the model performs a node evaluation as the aggregation of its children values. The type of aggregation operator for the concerned node in the template defines the operation. For example, in Figure 2 a mean calculation is used to evaluate the node "Consistent world" from the evaluations of "World is going on with or without the character" and "Natural mechanics, real formulas".

Our choice of aggregation operators was based on the study in [4] where a methodology for choosing an operator of aggregation corresponding to the behavior wished for the template is described.

Several weighted aggregation operators are implemented in the proposed model: four constant attitude operators (minimum, maximum, geometrical mean, arithmetical mean) and a variable attitude operator (symmetrical sum, [11]). This choice provides a very large spectrum of possible aggregations in order to express very different behaviors.

For example, in Figure 2, the aggregation "MAX" used for the evaluation of "Realism" has an optimistic behavior by taking at least the highest between the evaluations of "Graphical quality" and "Consistent world". Otherwise, the behavior of the aggregation "MEAN" has a compensatory effect between the leaves "World is going on with or without the character" and "Natural mechanics,

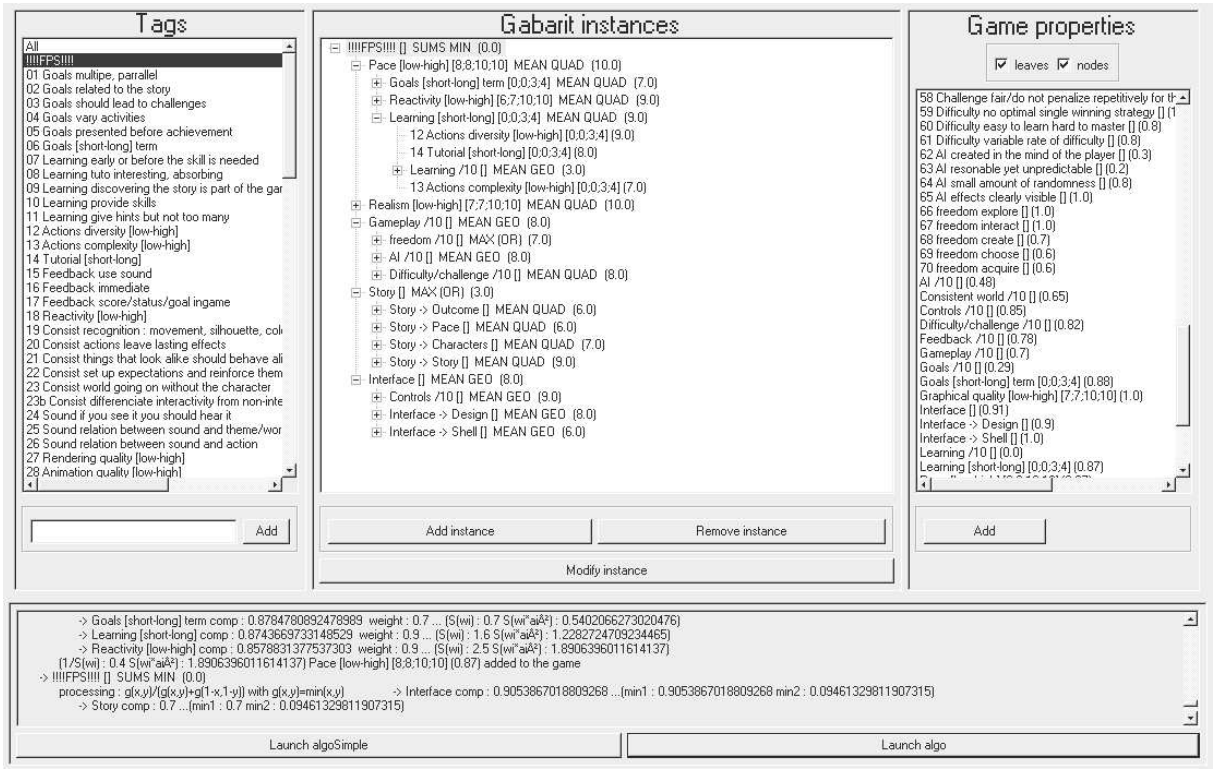

Figure 4: Screenshot of the evaluator prototype.

real formulas": a low evaluation of one may be counterbalanced by the other.

The weight assignment is a mean to take into account the importance of each criterion, thus our model can adapt even generic criteria to the type of the game.

4 Preliminary results

In order to validate our model, we created a template prototype for FPS video games and implemented a game evaluator to test this template.

4.1 Game evaluator prototype

The game evaluator prototype illustrated on Figure 4 contains a graphic user interface divided into four parts. The left part is used to define all the criteria of the template, in other words all linguistic pieces of information present in the template have to be set in this part. The center part provides the functionalities to create the template itself: one can add or remove nodes and leaves, choose the aggregation operators and define weights and fuzzy sets. The right part contains all the properties of the game to test. They are evaluated by players if they refer to leaves of the template, otherwise they are deduced automatically by the evaluator. Therefore, during the process, this part is progressively enriched with deduced properties until the root of the template is reached. The last part, on the bottom of the interface, is used to launch the evaluation process and visualise debugging information.

4.2 Experimental results

4.2.1 The FPS template

A FPS template was created and built into the evaluator. Its structure was adapted from the hierarchical game heuristic compilation presented in Section 2.2 in order to fit the characteristics of the FPS type. For example, the categories *Game story* and *Game Interface* were kept unchanged whereas the categories *Gameplay* and *Game mechanics* were modified or reorganized into new ones. A par-

Table 1: Video games assessment obtained from the template-based proposed approach for 3 FPS games: (left) "Bioshock" (Take 2 Interactive, 2007), (middle) "FarCry" (Ubisoft, 2004) (right) "Half-Life 2: Episode 2" (Valve, 2007).

Bioshock		FarCry		Half-Life 2: Episode 2	
Criteria	**Evaluation**	**Criteria**	**Evaluation**	**Criteria**	**Evaluation**
Gameplay	5.7	Gameplay	8.0	Gameplay	7.0
Interface	7.3	Interface	6.7	Interface	9.1
High pace	9.0	High pace	8.9	High pace	8.7
Realism	7.0	Realism	7.9	Realism	9.1
Story	7.3	Story	6.0	Story	7.0
Global	**8.5**	**Global**	**8.4**	**Global**	**8.9**

tial view of this template is presented in the center part of Figure 4.

4.2.2 Obtained results

Tests through interviews of experienced players were performed for three different games: "Bioshock" (Take 2 Interactive, 2007), "FarCry" (Ubisoft, 2004) and "Half-Life 2: Episode 2" (Valve, 2007). Players were asked to evaluate the elementary properties defined in our FPS template. These information were fed into the game evaluator that then computed a global assessment about the quality of the game.

The obtained results are presented in Table 1; for each game, the global evaluation is detailed by five intermediate marks: "Gameplay", "Interface", "High pace", "Realism" and "Story". They represent the last series of nodes leading directly to the root of the FPS template.

These three games received very good criticisms by video games journalists. The average rating is over 8.5 for these three games [7]. Our results perfectly reflect this general sentiment as the evaluator returned scores between 8 and 9.

The evaluation of "High pace" and "Realism" for the three cases is particularly high. We consider these aspects determinant in the quality of FPS games and this result matches our expectations. The scores for "Gameplay", "Interface" and "Story" are variable. They reflect positive and negative aspects of these games. For example, the "Story" score for Bioshock overtaking the one for FarCry is consistent with specialists assessments [7].

5 Conclusion and perspectives

The model proposed in this paper for evaluating video games proves to be a powerful tool to assess the quality of video games: preliminary tests on FPS games lead to results consistent with expert opinions.

Furthermore, the system offers an interpretable global quality, that can be followed back at several detail levels, thanks to its representation into a hierarchical structure. The use of linguistic variables makes the constraints and internal nodes understandable. It follows that every intermediate result of the evaluation process can be interpretated. Moreover, its advantages also comprise its flexibility, due the possibility of modifying e.g. the relative importance of the criteria, through the weighting coefficients or through the aggregation operators. Thus, video game experts have a complete control on the system as they have the opportunity to justify the overall process or on the contrary modify it.

The quality of the obtained results depend on the built template, which in particular requires the determination of its global structure, aggregation operators and possibly weighting coefficients. This construction step is a complex one; perspectives of this work include the development of methods to auto-

matically build the templates. In particular they could be based on a preliminary machine learning step, aiming at extracting common or typical properties of good and bad video games respectively.

It can be noticed that the prototype built for FPS games already provides a firm ground for building templates for other game types. Nevertheless, further studies should be conducted with the help of video game experts and users in order to validate and complete our criteria classification. Moreover, one can imagine building a template specialized e.g. on ergonomic criteria and compare the results with existing ergonomic evaluations.

Another perspective aims at interfacing the proposed evaluator with physiological measurement systems: the idea is to replace some scores directly provided by the player answers with information deduced from her physiological state. For example, criteria involving the player interest in the story could be derived from heart rate measures recorded at appropriate moments of the game. This approach should make it possible to further reduce the quantity of questions the player has to answer and possibly gain objectivity.

References

[1] E. Brown and P. Cairns. A grounded investigation of game immersion. In *Proceedings of the International Conference on Human Factors in Computing Systems*, pages 1297 – 1300. ACM New York, NY, USA, 2004.

[2] E. Collain. Technique des gabarits. Technical report, Thomson-CSF, 1995.

[3] H. Desurvire, M. Caplan, and J. Toth. Using heuristics to evaluate the playability of games. In *Proceedings of the International Conference on Human Factors in Computing Systems*, pages 1509 – 1512. ACM New York, NY, USA, 2004.

[4] V. Eude. *Modélisation spatio-temporelle floue pour la reconnaissance d'activités militaires*. PhD thesis, Université Paris 6, 1998.

[5] N. Falstein and H. Barwood. The 400 project. Available at http://theinspiracy.com/400_project.htm.

[6] M. Federoff. Heuristics and usability guidelines for the creation and evaluation of fun in video games. Master's thesis, University Graduate School of Indiana University, 2002.

[7] Gamerankings. 2007. Available at http://www.gamerankings.com/.

[8] Y. Isomoto, K. Yoshine, H. Nakatani, and N. Ishii. Data model and fuzzy information retrieval for scenic image database: theoretical extension of a traditional crisp model to fuzzy model. *Fuzzy information engineering*, pages 283–289, 1997.

[9] L. Mouillet. *Modélisation, reconnaissance et apprentissage de scénarios de conflits éthno-politiques*. PhD thesis, Université Paris 6, 2005.

[10] L. Mouillet, B. Bouchon-Meunier, and E. Collain. Automated identification of political conflicts with a scenario recognition technique. In *Proceedings of IPMU*, volume 3, pages 1609–1616, July 2004.

[11] W. Silvert. Symmetric summation: a class of operations on fuzzy sets. *IEEE Transactions on Systems, Man and Cybernetics*, 9:659–667, 1979.

[12] P. Sweetser and P. Wyeth. Gameflow: A model for evaluating player enjoyment in games. *ACM Computers in Entertainment*, 3(3), 2005.

[13] P. Voderer, T. Hartmann, and C. Klimmt. Explaining the enjoyment of playing video games: the role of competition. In *ACM International Conference Proceedings Series*, volume 38, 2003.