Animated movie genre detection using symbolic fusion of text and image descriptors

Gregory Païs, Patrick Lambert, Françoise Deloule, Daniel Beauchêne, Bogdan Ionescu

Animated Movie Genre Detection using Symbolic Fusion of Text and Image Descriptors

Gregory Païs[(1)], Patrick Lambert[(1)], Daniel Beauchêne[(1)], Françoise Deloule[(1)], Bogdan Ionescu[(2,1)]
[(1)] LISTIC - University of Savoie - 74940, Annecy-le-Vieux, France
{gregory.païs, patrick.lambert, daniel.beauchêne, françoise.deloule}@univ-savoie.fr
[(2)] LAPI - University Politehnica of Bucharest, Romania - bionescu@alpha.imag.pub.ro

Abstract

This paper addresses the automatic movie genre classification in the specific case of animated movies. Two types of information are used. The first one are movie synopsis. For each genre, a symbolic representation of a thematic intensity is extracted from synopsis. Addressed visually, movie content is described with symbolic representations of different mid-level color and activity features. A fusion between the text and image descriptions is performed using a set of symbolic rules conveying human expertise. The approach is tested on a set of 107 animated movies in order to estimate their "drama" character. It is observed that the text-image fusion achieves a $precision$ *up to* 78% *and a* $recall$ *of* 44%.

1. Introduction

Nowadays, very large collections of images and videos are continuously created in many domains: medicine, geographic, surveillance, entertainment, etc. The quantity and the diversity of these collections are gigantic and managing such collections in an efficient way is very difficult. So the need for building software tools helping in this management is crucial, specifically in the case of videos where the size, the richness and the variety of data is considerable. Generally, to design such tools, the first step which is required is the indexation of the documents. Classically, these index are mainly textual index (keywords) provided manually in many cases.

However, in the two last decades, a lot of work has been dedicated to automatic indexing. Among all these works, automatic cataloging of videos or movies into predefined semantic categories (i.e genres) is a very interesting challenge. This cataloging can be performed either globally, like classifying videos into one of several main genres (cartoons, music, news, sports) or into some sub-genres, e.g. identifying specific types of sports (football, hockey, etc.), movies (drama, thriller, etc.) either locally and thus focusing on classifying segments of video such as retrieving concepts: outdoor vs. indoor, violence, action.

Different approaches have been proposed. A state-of-the art is available in [1]. [8] discusses an uni-modal (image) approach and testes the prospective potential of motion information to cartoon classification. However, results are obtained on a very limited data set (8 cartoon and 20 non cartoon sequences). A two-modal (image and audio) approach is proposed in [3] and cartoon classification is performed using a multilayered perceptron with both visual (brightness, saturation, color hue, edge information, motion) and audio descriptors (MFCC descriptors). Tests were performed on a bit larger database containing 100 sequences (20 sequences of each genre: cartoons, commercials, music, news and sports) and classification accuracy is around 90%.

Another example is the approach in [10] which uses Bag-of-Visual-Words with shot classes as vocabularies. The classification is performed into four genres (Action, Comedy, Drama and Horror) on 1239 annotated trailers. Obtained accuracy is between 60% and 80%. [7] proposes a truly multi-modal approach which combines several types of content descriptors. Features are extracted from four informative sources, which include visual-perceptual information (color, texture and motion), structural information (shot length, shot distribution, shot rhythm, shot clusters duration and saturation), cognitive information (face properties, such as number, positions and dimensions) and aural information (transcribed text, sound characteristics). These pieces of information are used to train a parallel neural network system and provide a maximum accuracy rate up to 95% in distinguish between seven video genres (including cartoons): football, cartoons, music, weather forecast, newscast, talk shows and commercials. However, these techniques, in general, are not focusing on the retrieval of animated content. They are limited to use "all purpose" content descriptors which work well with all genres, but not

specifically with animated content.

The work which is proposed in this paper is related to automatic cataloging, with two main specificities. First, the source data are original as they are composed of artistic animated movies of the "Animated Film International Festival" (CITIA). This festival, which states each year in Annecy from 1960, is one of the major events in the worldwide animated movie entertainment, providing a video movie library with more than 30.000 movie titles. Animated movies from CITIA are different from classic animation movies (i.e., cartoons) or from natural movies in many respects (short duration, broad variability content, artistic effects, etc.). Second, the categorization which is proposed is a sub-genre categorization within the animated movie genre. The specific case of sub-genre cataloging is more difficult because it makes reference to small differences, which are difficult to detect in an automatic way, and is generally related to some semantic information.

To cope with these difficulties, the solution which is proposed in this work is to use jointly information extracted from images and textual information provided by synopses (at the moment, audio is not used as the relation with movie genre is more complex). The main difficulty in using jointly image and textual information is their very different semantic levels. Generally, image features are low-level features as textual information has a high semantic level. So to aggregate these very different information, we propose to use symbolic descriptions and symbolic rules. On one hand, symbolic description is an easy way to get a common representation of very different type of information. On an other hand, symbolic rules is an elegant and efficient manner to covey human expertise. (In [4], such a symbolic description, but limited to image content, has already been used in the case of animated movie characterization).

The paper is organized as follows. Section 2 (respectively 3) details how synopses (respectively images) are used to estimate movie genre. The fusion between textual and image characterization are proposed in Section 4. Section 5 presents the conclusions and proposes future works.

2. Text-based genre characterization

When the movies are registered, a set of textual information are required: year, country, director, duration, title, synopsis, etc. Among all these textual information, synopsis is a crucial one as it contains a very rich information about movie content, and particularly about movie genre. Indeed, a natural idea is to consider that the vocabulary used in the synopsis has a high probability to contain words directly related to the movie genre. The following synopsis example confirms this assumption for the movie "Romeo and Juliette" which genre is "Drama":

"A travelling group of monsters, dragons and ugly creatures, in town for just one day, will perform the famous love tragedy by William Shakespeare for the movie";

However this natural assumption is not always so clear, as it can be seen in the following synopsis (movie: "Circuit Marine", genre: "Adventure"):

"To be eaten or not to be eaten? That is the question!";

It can be noted that the genre is also an information which is required when registration is performed. However, in many cases, genre is not filled in, or given in an imprecise way. For instance "Artistic" is a very common declared genre, which of course is not very relevant for the classification. So the declared genre will only be used to build and test the proposed strategy which consists in estimating genre through synopsis analysis.

The basic principle is based on the definition of a *lexical intensity* for each synopsis and regarding each genre. Then, the estimated genre will be the one corresponding to the greatest *lexical intensity*. For a specific genre, *lexical intensity* is defined as the proportion of words within the synopsis belonging to the thematic dictionary of the considered genre. This approach requires thematic dictionary for each genre. As such dictionaries do not exist for animated movie genres, the first step consists in the definition of these dictionaries.

2.1 Thematic dictionary

First, we have selected 5.804 synopses associated to movies for which the genre has been correctly declared. These synopses are regarding as a training corpus. This corpus is lemmatized, i.e. each textual form is replaced by its lemma (verbs with infinitive, plural nouns with singular, adjectives with singular masculine form) and non relevant terms as articles, etc. are eliminated. For each genre g, we first defined a sub-corpus which is composed of all the lemmas found in a synopsis for which the declared genre is g. Then, a statistical procedure is used. Let N denotes the total number of different lemmas in the all corpus and N_g the total number of different lemmas corresponding to a genre g. For a given term m, $N(m)$ denotes the number of times this term appears in the whole corpus and $N_g(m)$ the number of times it appears in the sub-corpus corresponding to the genre g. The specificity index of m in g is defined by :

$$I_g^{spe}(m) = \frac{N_g(m)}{N(m)} \cdot \frac{N}{N_g} \quad (1)$$

when $I_g^{spe}(m)$ is greater than 1, the term m is more frequent than other terms for the genre is g. So this term can be used in the thematic dictionary describing the genre g.

Thus, we obtain an index of specificity for each lemmatized term of the corpus and for each genre. Then for each studied genre, we only keep in storage the terms

whose specificity index is higher than 1.5 (empirically determined). At this stage, we introduce a manual step to eliminate the terms which are not really related to the considered genre. The remaining terms are labeled as "seeds" which are used to complete the thematic dictionary. This is realized by using a synonym dictionary (CRISCO, see `http://dico.isc.cnrs.fr/dico/en/chercher`). Starting from the "seeds", we retain the synonyms (at level 1 and 2). A last manual filtering is necessary to eliminate the "noise" due to polysemia. The list of these terms constitutes the thematic lexicon of each genre. For example, in the case of "Drama", the specificity index allows the selection of 1900 terms. After the manual filtering, it remains a list of 101 terms. After looking for the synonyms and a last filtering, the final "Drama" thematic dictionary is composed of nearly 800 terms.

2.2 Thematic intensities

Thanks to these thematic lexicons, it is possible, for each synopsis, to compute the thematic intensity on each genre. We define the *thematic intensity*, $I_{th}(sy, g)$ of a synopsis sy for a genre g as the ratio between the number of lemmas which belong to the thematic lexicon of the genre and the total number of lemmas of the synopsis:

$$I_{th}(sy, g) = \frac{|T_{sy} \bigcap T_g|}{|T_{sy}|} \quad (2)$$

where T_{sy} is the lemma set of synopses sy and T_g is the thematic dictionary for the genre g, $|X|$ denoting the cardinality of set X.

To check the discriminating capacity of this indicator, we have computed the "Drama" *Thematic Intensity* on all the 5804 synopses. Then, for each declared genre, we have computed the average of the "Drama" *Thematic Intensities* $\overline{I_{Drama}}$. We obtain the results presented in Table 1.

Table 1. "Drama" thematic intensity (%) vs. declared genre.

Declared genre	$\overline{I_{Drama}}$	Declared genre	$\overline{I_{Drama}}$
Adventures	4.30	Fantasy	4.43
Comedy	5.93	Humour	5.26
Tale	3.92	**Black humor**	**9.52**
Documentary	3.09	**Thriller**	**9.16**
Drama	**8.28**	Propaganda	4.66
Instructive	4.56	Provocative	5.25
Epic	4.19	Advertising	1.22
Erotic	3.44	Satire	6.16
Science-fiction	6.03		

As expected, it can be noted, that "Drama" has a high average (8.28). But "Black Humor" and "Thriller" have also averages significantly higher than the other genres. This last result is not surprising as these 2 genres may be close to "Drama" genre. We use these results to define a simple strategy for classifying synopses. The rule is the following:

IF $(I_{th}(sy, Drama) \geq 0.08)$ **THEN** $(sy \in$ "Drama"),

By comparing the results obtained with this rule with the declared genre, considered as ground-truth, we obtain the confusion matrix in Table 2 (where TN stands for *True Negative*, FP for *False Positive*, FN for *False Negative* and TP for *True Positive* - for an overview of retrieval statistics see `http://www-nlp.stanford.edu/IR-book/`).

Table 2. Confusion Matrix on "Drama".

$Declared/Estimated$	$Not\ Drama$	$Drama$
$Not\ Drama$	2892 (TN)	614 (FP)
$Drama$	363 (FN)	194 (TP)

With this classification, *precision* (P) is 24%, *recall* (R) is 34% and F_{score} is 28%. These results are rather bad. There are two main causes. On one hand a dramatic atmosphere is not always discernable in the synopsis vocabulary as it may require a high level comprehension. On the other hand, movie synopses may contained words related to "Drama" when the declared genre is not Drama. This declared genre may be "Thriller" or "Black Humour" as it has been already noted (see table 1), this genres having frequently a dramatic atmosphere.

Tests have also been conducted on other genres. With "Thriller", we got F_{score} = 30% which similar to "Drama". With Humour, performance is worse (F_{score} = 16%).

These results can be improved by taking into account the synopsis length. The basic principle is to enforce the thematic intensity when the synopsis is long. Indeed, in a long synopsis, the weight of relevant lemmas may be weakened by other non relevant words. The fusion between $Length$ and *Thematic Intensity* is performed in a symbolic way. First, the numeric values of $Length$ and *Thematic Intensity* are transformed in symbols using a fuzzification step. Five symbolic variables are defined for *Thematic Intensity* (Very Short (VS), Short (S), Medium (M), Long (L) and Very Long (VL)) and five symbols are defined for $Length$ (Very Low (VL), Low (L), Medium (M), High (H), Very High (VH)). Then a symbolic fusion is realized according to a set of symbolic IF-THEN rules which are summarized in Figure 1 and which define three symbols for the improved *Thematic Intensity* (Low (L), Medium (M), High (H)). The following example illustrates the way this table has to be read:

IF $(Drama\ intensity$ **is** $Medium)$ **AND** $(Synopsis\ Length$ **is** $High)$ **THEN** $(Drama$ **is** $Medium)$,

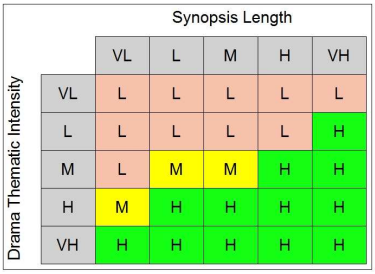

Figure 1. Fuzzy rules defining "Drama".

Finally a fuzzy inference is used to determine the membership degree of each of the three output symbols. The fuzzy inference is realized with product as t-norm and Lukasiewicz *t-conorm* [6]. From this measure, a crisp measurement is obtained by taking the output symbol corresponding to the greatest membership degree and by accepting the corresponding genre if the output symbol is Medium or High.

We test this solution always for the "Drama" genre estimation on the 5.804 synopses of the database and compare the results obtained with the declared genre. We achieve

$$P = 23\%, \quad R = 44\%, \quad F_{score} = 30\%.$$

Recall value is clearly increased (34% before adding synopsis length) when Precision value remains the same (24% before adding synopsis length). Improvement is also obtained on "Thriller" estimation ($F_{score} = 40\%$ instead of 30%). However, with "Humour" performance is still reduced ($F_{score} = 17\%$).

Tests have also been conducted on other genres. With "Thriller", we obtained $F_{score} = 30\%$ which similar to "Drama". With "Humour" performance is even worse, $F_{score} = 16\%$.

In the specific case of "Drama", as confusion may occur with "Thriller" and "Black Humour", an other way to improve performances consists in enhancing "Drama" estimation when the synopsis contains "Drama" vocabulary without containing "Thriller" or "Black Humour" vocabulary. This is obtain thanks to an another set of IF-THEN symbolic rules (for brevity reasons we won't provide these details). It is important to note that all these rules as well as the symbolic transformations represent the introduction of a human expertise and plays a similar role as the ground truth used in machine learning approaches. Of course the design of these symbolic transformations or rules has an influence on the performance.

3. Image-based genre characterization

Classically, image analysis provide low or medium-level descriptions of the image or video content compared to the high semantic level of descriptors extracted from textual information. In the case of videos, these descriptors are related to color, shape, texture or motion. Combining a lot of low-level descriptors and introducing human expertise is a way to improve significantly the description level of image descriptors.

In the proposed work, low-level features which are used have two main specificities.

- First, they give a global characterization of each movie. This type of description is relevant for animation movies which duration is small (typically 10 minutes) and would probably fails in the case of long duration movies or videos.
- Second, they are related to the movie color distribution and to the movie activity: the movie color distribution provides us with detailed information regarding the movies artistry content while the movie activity provides us with information on the movie action content.

The solution which is proposed here consists in using some numerical features related to color and activity and then in giving them a symbolic description. It is limited to the characterization of "Drama" character.

3.1 Low-level image descriptors

3.1.1 Color descriptors

After a temporal segmentation determining the different shots, a reduced color palette (216 colors - see `http://www.visibone.com/colorlab`) is used to get the global weighted color histogram $h_m(c)$ of each movie m [4] (c denotes the color index). It can be noted that in this palette each color is defined by three symbols characterizing its hue, saturation and intensity. For instance, the color (R=204, G=0, B=102) will be symbolically described by the three symbols: "Pink, Hard, Dark".

In a second step, several color features are extracted from this histograms: the warm color ratio f_{warm}, the dark color ratio f_{dark} and the color variation ratio f_{var}. The warm color ratio is defined as the proportion of warm colors in h_m. A warm color is a color which symbolic description contains one word of the set Γ_{warm} = {"Yellow", "Orange", "Red", "Yellow Orange", "Red Orange", "Red Violet", "Magenta", "Pink" or "Spring"}. Then, f_{warm} is defined by:

$$f_{warm} = \sum_{c=1}^{216} h_m(c)|_{Name(c)\in\Gamma_{warm}} \qquad (3)$$

As for the textual features, a fuzzy description of f_{warm} is defined with three symbols Low, $Medium$ an $High$. Other features are obtained in a similar way (see [4] for more details).

3.1.2 Activity descriptor

There are many ways to measure the activity within a movie. In [9] the activity is directly related to the density of shot transitions, the content variation within a shot being neglected. This approach uses a relatively confirmed assumption that, in general, action content is related to a high frequency of shot changes. On the contrary, in [2] activity is an intra-shot measure, which is also a reasonable assumption. So, based on the work in [5], the solution we use here is a mix between these two strategies. The basic idea is to accumulate the differences between successive frames until the cumulated difference exceeds a pre-defined threshold (empirically determined). Then, a key-frame is selected, and the accumulator is reset, and the mechanism is iterated. It is clear that the cumulated difference is both feed by transitions and by intra-shot variations. The frame difference which is used is a bloc-color difference associated to a motion compensation.

Then, the activity is measured through the frequency of selected key-frames in the following way:

- First, for each key frame we compute the time distance (in seconds) to the following key frame. This distance is denoted $Delta(i)$, i being the key frame index;
- Then, we compute the average all theses distances $\overline{Delta}$;
- Finally, we define the movie global activity by:

$$f_{activity} = \frac{4}{F_R} \cdot \frac{1}{\overline{Delta}} \quad (4)$$

where F_R is the frame rate. The coefficient $4/F_R$ aims to provide a feature normalized within the range $[0; 1]$.

As for other features, a fuzzy description of $f_{activity}$ is defined with three symbols Low, Medium and High.

3.2 Information fusion for Drama genre

To evaluate the "Drama" character of a movie thanks to its image content, the different symbols are fused according different symbolic fusions. Figure 2 presents the organization of these fusions. Using the symbolic representation of the different image features previously defined and denoted, three different symbolic fusions are designed to define three symbols: the Coldness, the Monotony and the Uniformity of a movie. Then, these three symbols are fused to define the Movie "Drama" Character.

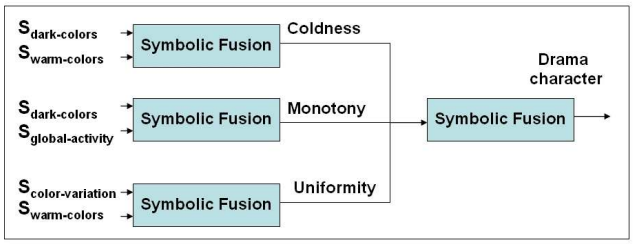

Figure 2. Fusion example of symbolic image descriptors.

As in the previous fusion, these different symbolic fusion are performed using set of symbolic IF-THEN rules. Due to lack of space, these rules are not presented here. We only limit the presentation of the situation which gives a High "Drama" Character (H and M denote High and Medium):

Drama Character **is** H **IF**
$\{$(Cold. **is** M) **AND** (Monot. **is** H) **AND** (Unif. **is** H)$\}$
OR $\{$(Cold. **is** H) **AND** (Monot. **is** M) **AND** (Unif. **is** H)$\}$
OR $\{$(Cold. **is** H) **AND** (Monot. **is** H) **AND** (Unif. **is** M)$\}$

3.3 Performances

This approach has been tested on a set of 107 movies in the case of "Drama" Characterization. The obtained results are presented in Table 3.

Table 3. Drama genre estimation with image features.

	Cold.	Monot.	Unif.	Fusion
$precison$	28%	20%	28%	39%
$recall$	81%	69%	85%	56%
F_{score}	42%	31%	42%	46%

It can be noted that the fusion between the three symbolic characterization obtained from images increases the $precision$ while $recall$ decreases. Globally, F_{score} increases which demonstrates the interest of using jointly different characterization. On the same set of 107 movies, using only synopsis as it has been described in Section 2, the results were:

$$P = 43\%, \quad R = 81\%, \quad F_{score} = 46\%.$$

4. Fusion between text and image information

The final step consists in using both synopses and images. The strategy, following the previous fusion mechanisms, is based on a symbolic fusion between the two symbolic characterizations respectively obtained from synopses and images. The rules are summarized in Figure 3 where "Drama" Character is defined by three symbols Low (L), Medium (M) and High (H).

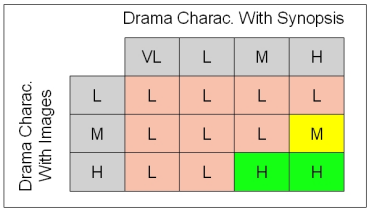

Figure 3. Rules for synopsis / image fusion.

It can be noted that, as synopses provides better marginal results than images, rules put a little bit more emphasis on characterization obtained from textual information. On the same set of 107 movies, we obtained the results given in Table 4.

Table 4. Drama genre estimation with synopsis and images.

	Synop.	Image	Synop. & Image
$precison$	43%	28%	78%
$recall$	81%	85%	44%
F_{score}	56%	42%	56%

Fusion between textual and image information increases the $precision$ in a very significative way. However, this goes with a decrease of $recall$. It means that, in the case of "Drama", using both synopsis and images allows to get a relevant response when estimated genre is declared to be "Drama". Unfortunately, nearly half of the "Drama" movies are not detected with this measure.

Using the same features (the membership degrees) to get a relevant comparison, the proposed method is compared to two supervised classifiers: the Multi-Layer Perceptron (MLP) and a Support Vector Machine (SVM) classifier with a Sequential Minimal Optimisation using a cubic polynomial kernel. A cross validation (with a factor 3) has been realized and we got the following results:

$$SVM = 47\%, \quad R = 68\%, \quad F_{score} = 56\%,$$
$$MLP = 44\%, \quad R = 44\%, \quad F_{score} = 44\%,$$

The symbolic text-image fusion provide a better $precision$ but a lower $recall$ compared to SVM approach, while global performances are similar ($F_{score} = 56\%$). MLP is less accurate.

5. Conclusions

This work is a first study for estimating automatically the genre of animated movies. It is based on a symbolic fusion between textual information (extracted from synopsis) and color and activity information (extracted from images). It has been illustrated on the "Drama" character estimation. Results are encouraging, specifically concerning $precision$. Thanks to the symbolic approach, the proposed approach is able to incorporate expert knowledge. Many others works have to be engaged: the approach has to applied on other genres, audio information could be added, the test database has to be extended, image features can be enriched...

References

[1] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C:Applications and Reviews, 38(3)*, pages 416–430, 2008.

[2] W. Farag and H. Abdel-Wahab. Adaptive key frames selection algorithms for summarizing video data. Technical report, Norfolk, VA, USA, 2001.

[3] R. Glasberg, A. Samour, K. Elazouzi, and T. Sikora. Cartoon-recognition using video and audio-descriptors. In *EUSIPCO*, Antalya, Turkey, Sept. 2005.

[4] B. E. Ionescu, D. Coquin, P. Lambert, and V. Buzuloiu. Fuzzy semantic action and color characterization of animation movies in the video indexing task context. In *AMR 4th International Workshop on Adaptive Multimedia Retrieval*, volume 4398, pages 119–135. Springer-Verlag, 2007.

[5] T. Lu and P. N. Suganthan. An accumulation algorithm for video shot boundary detection. *Multimedia Tools Appl.*, 22(1):89–106, 2004.

[6] G. Mauris, E. Benoit, and L. Foulloy. The aggregation of complementary information via fuzzy sensors. *Measurement*, 17(4):235–249, 1996.

[7] M. Montagnuolo and A. Messina. Parallel neural networks for multimodal video genre classification. *Multimedia Tools Appl.*, 41(1):125–159, 2009.

[8] M. Roach, J. Mason, and M. Pawlewski. Motion-based classification of cartoons. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 435–444, Hong-Kong, 2001.

[9] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X.-Q. Wu, and S. Li. Automatic video genre categorization using hierarchical svm. In *IEEE Int. Conf. on Image Processing*, pages 2905–2908, 2006.

[10] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg. Movie genre classification via scene categorization. In *Proceedings of the international conference on Multimedia*, MM '10, pages 747–750, New York, NY, USA, 2010. ACM.