De la détection d'évènements sonores violents par SVM dans les films

Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, Patrick Gros

De la détection d'évènements sonores violents par SVM dans les films *

Cédric Penet[1,2] Claire-Hélène Demarty[2] Guillaume Gravier[1] Patrick Gros[1]

[1] INRIA Rennes & IRISA/CNRS
[2] Technicolor Rennes

1 avenue de Belle Fontaine, 35510 Cesson-Sévigné
Cedric.Penet@technicolor.com

Résumé

*Cet article étudie le comportement d'une approche classique, à l'état de l'art, pour la détection d'événements sonores par machines à vecteurs supports, appliquée à la détection d'événements violents dans les films. Les événements sonores considérés, liés à la présence de violence, sont les **Cris**, les **Coups de feu** et les **Explosions**. Nous montrons que, contrairement aux résultats d'autres études, l'approche état de l'art ne donne pas de bons résultats sur cette tâche. Une étude sur la répartition des échantillons en sous-ensembles dans un protocole de validation croisée permet d'expliquer ces résultats et met en évidence un problème de généralisation, dû au polymorphisme des classes considérées. Ce polymorphisme est démontré par un calcul de divergence entre les échantillons de la base de test et ceux de la base d'apprentissage.*

Mots Clef

Détection d'évènements, audio, SVM, généralisation, validation croisée.

Abstract

*This article studies the behaviour of a state-of-the-art support vector machine audio event detection approach, applied to violent event detection in movies. The events we are trying to detect are **screams**, **gunshots**, **explosions**. Contrary to others studies, we show that the state-of-the-art approach does not lead to good results on this task. A study on the repartition of samples into subsets in a cross validation protocol helps explain those results and highlights a generalisation problem due to a polymorphism of considered classes. This polymorphism is demonstrated by the computation the divergence between the samples of the test database and the training database.*

Keywords

Event detection, audio, SVM, generalisation, cross-validation.

*Travaux partiellement financés par le projet Quaero.

1 Introduction

L'indexation et l'analyse de contenus multimédias sont devenues des éléments essentiels de la chaîne de production de la télévision. Cette dernière possède différents programmes, dont certains récurrents (comme le journal télévisé), qui sont souvent archivés après diffusion pour utilisation ultérieure, par exemple dans le cadre de la vidéo à la demande (VOD) ou de la télévision de rattrapage. Cet archivage doit faire l'objet d'une annotation qui va permettre de renseigner sur le contenu des documents. Ce travail est d'autant plus fastidieux qu'il y a d'informations à annoter. C'est pourquoi il est nécessaire d'automatiser au maximum le processus.

La détection automatique d'évènements dans des vidéos découle directement de ce besoin et est largement étudiée. Dans [17], les auteurs font une revue de l'état de l'art sur la détection d'évènements dans les flux multimédias, donnant de nombreuses références sur le sujet. Les évènements considérés peuvent être de natures différentes, comme des portes qui s'ouvrent, des personnages qui parlent, de la violence, etc. Notre étude s'inscrit dans un projet de détection de la violence dans des films. L'objectif est de pouvoir présenter à un utilisateur de VOD les scènes les plus violentes d'un film dans le but de l'aider à faire un choix. A cette fin, nous nous intéressons dans cet article à la détection d'évènements audio violents dans les films.

Si la détection d'évènements au sens général du terme est largement étudiée, celle de la violence dans des films l'est en revanche beaucoup moins. Ceci peut s'expliquer par le fait qu'action et violence sont souvent confondues comme cela apparaît dans la littérature. Par exemple, dans [9], les auteurs décrivent les scènes d'action comme devant être excitantes et maintenir en haleine le public, ne pas lui laisser le temps de se reposer. Ce type de scènes peut parfois être assimilé à de la violence. Une poursuite en voiture, par exemple, ne peut pas être systématiquement classée dans la catégorie scène violente, même si elle répond à tous les critères de la scène d'action. Dans [16], les auteurs définissent quatre classes sémantiques : trois classes d'action (**Running, Chasing and Object Tracking**, **Figthing** and **Explosions and Crashing**) et une classe **Panorama** défi-

nie par de faibles mouvements dans l'image. Sur les trois classes d'action, seules les deux dernières peuvent être objectivement classées comme violentes alors que la première ne l'est pas toujours. De même, dans [3, 2], les auteurs se sont concentrés sur les combats personnes contre personnes, l'un les assimilant à de l'action, l'autre à de la violence.

Nous pensons que la violence n'est pas toujours liée à de l'action, et inversement. Certaines scènes violentes peuvent ainsi répondre à des critères pouvant être assimilés à de la non-action, dans le cas par exemple de violence contextuelle. Des scènes présentant une menace de mort violente pour l'un des personnages peuvent ne pas être des scènes d'action et ne répondre à aucun des critères correspondants mais être très violentes psychologiquement et contextuellement parlant. Par exemple, dans [11, 13], les auteurs ont essayé de reconnaître des motifs dans la dynamique du son dans des films d'horreur. Ces motifs sont parfois annonciateurs de scènes violentes et/ou angoissantes. Dans [15], les auteurs ont fait la même chose en se focalisant cette fois-ci sur l'intensification de la lumière dans la vidéo.

Il est important de noter que le son joue un rôle à part entière dans le caractère violent ou non d'une scène. En effet, une scène violente sera souvent accompagnée d'une certaine ambiance sonore ou de sons caractéristiques du type de scène violente que l'on considère. C'est pourquoi la modalité sonore apparaît comme une modalité importante pour la détection des scènes violentes. Dans le cadre de notre article, nous nous sommes concentrés sur cette modalité uniquement.

Parmi les travaux se rapprochant le plus de notre cas d'utilisation, se trouvent les travaux de [5, 6, 14, 7]. Dans ces articles, les auteurs essayent de détecter la violence à travers les sous-classes suivantes : **musique, parole, coups de feu, combats, cris, et des classes environnementales**, en utilisant majoritairement le son (sauf [7], qui utilise l'image en plus du son).

Bien qu'étant de natures différentes, ces travaux ont tous en commun le fait qu'ils utilisent une petite base de d'apprentissage et/ou de test. Le nombre d'échantillons est généralement très restreint (entre 50 et 500 en général) et provient de peu de films (en général 4 ou 5). Par exemple, dans [9], les auteurs travaillent sur quatre films et cinquante sept échantillons. La diversité du problème, le nombre de films présentant de la violence, la difficulté de définir exactement les scènes violentes comparés au peu d'exemples utilisés pour construire les modèles nous amènent à nous poser des questions sur les propriétés de généralisation de ces systèmes.

Une autre remarque peut être faite sur la façon dont les données disponibles sont utilisées dans ces travaux. De manière générale, les données sont agrégées puis séparées en base de test et base d'apprentissage, de sorte que chacun des documents utilisés pour les construire ait des échantillons dans chacune des bases. Nous montrons que ceci pose un problème dans le sens où les données sur lesquelles le système sera utilisé seront des données totalement nouvelles. Nous pensons qu'en séparant les données en deux bases sans tenir compte de leur provenance, les expériences ne sont pas représentatives de la réalité.

Attributs	Dimension
Mel-Frequency Cepstral Coefficients (MFCC)	12
Énergie	1
Centroïd fréquentiel	2
Asymétrie spectrale	2
Platitude spectrale	3
TOTAL	**20**

TABLEAU 1 – Attributs extraits.

Dans cet article, le système que nous présentons est axé sur la détection de trois classes d'évènements audio violents qui sont : **Cris, Coup de feu et Explosions**. Il utilise des techniques niveau état de l'art. Nous utilisons une base de données de 16 films, ce qui est plus que pour les systèmes que l'on peut trouver dans la littérature et nous montrons que ce système classique mène vers un problème de généralisation. Nous calculons la divergence entre les échantillons de la base de test et de la base d'apprentissage pour tenter d'expliquer ce problème de généralisation.

La section 2 décrit brièvement le système utilisé, la section 3 détaille les spécificités du protocole expérimental utilisé, la section 4 présente les résultats et enfin, la section 5 est dédiée à une discussion pour tenter d'expliquer les résultats obtenus.

2 Présentation du système

La figure 1 présente le processus utilisé dans son ensemble. Il se décompose classiquement en deux grandes parties : une partie apprentissage et une partie inférence.

2.1 Segmentation du flux audio

Le flux audio des films est segmenté en fenêtres de 40 ms. Ne connaissant pas la durée moyenne des sons que nous essayons de détecter, nous choisissons de rester sur cette base de 40 ms afin d'éviter de trop agréger. De plus, la question de la validité de statistiques extraites sur une vingtaine ou même une centaine d'échantillons se pose.

2.2 Extraction d'attributs

Les attributs sont extraits directement sur les échantillons de 40 ms. Ce sont des attributs classiques largement utilisés dans la littérature. Le tableau 1 liste les attributs calculés. Les attributs sont ensuite normalisés entre -1 et 1, pour réduire les effets dus à des dynamiques trop différentes d'un attribut à l'autre. Au final, un vecteur de dimension $n = 20$ est extrait de chaque échantillon pour être donné au classifieur.

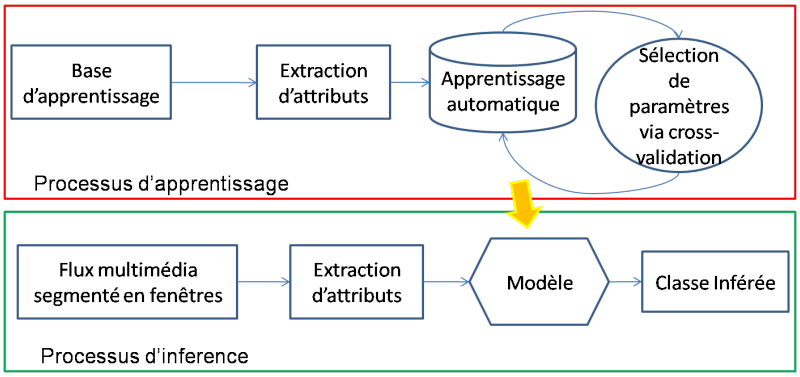

FIGURE 1 – Processus utilisé

2.3 Classification

Pour ce qui est de la classification, nous utilisons l'algorithme de machine à vecteur support (SVM) [1]. Nous entrainons un C-classifieur à vecteur support (C-SVC) avec un noyau de fonctions de base radiales (RBF). De même, ce choix a été guidé par la large utilisation et le succès des SVMs dans la littérature [1, 2, 12, 5].

3 Protocole Expérimental

Le problème que nous nous proposons de résoudre est un problème multi-classes comprenant quatre classes : **Cris (S), Coups de feu (G), Explosions (E) et Autres (O)**. La dernière classe est composée de tout ce qui n'appartient pas aux trois autres classes.

3.1 Base de données

Comme précisé dans l'introduction, notre base de données contient 16 films annotés manuellement. La base contient des films non violents comme *Le magicien d'Oz*, des films très violents comme *Léon* ou *Reservoir Dogs*, et des films "tout public" tels que *Harry Potter et L'ordre du Phénix* ou encore *Retour vers le futur*.
Concernant la répartition des échantillons de chaque classe dans la base de données, deux propriétés importantes sont à noter. Premièrement, les échantillons **O** correspondent à environ 98% de la base de données et sont donc fortement majoritaires. Deuxièmement, les films sont très inhomogènes entre eux. Certains vont posséder des échantillons des trois classes violentes (S, G et E) en grande quantité, comme les films très violents, tandis que d'autres vont n'en posséder que quelques uns seulement, et pour quelques classes seulement (ainsi, *Billy Elliot* ne va pas contenir d'explosions). Le problème que nous avons à résoudre est donc particulièrement difficile à appréhender, comprenant un cas de déséquilibre important entre les classes (une classe est largement majoritaire, même dans les films contenant les trois autres classes en quantité), conjugué à un problème de déséquilibre important entre les documents (les films ne contiennent pas tous tous les évènements considérés).

1. Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines, 2001. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

Classe	Nombre d'échantillons
Cris (S)	6 743
Coups de feu (G)	6 746
Explosions (E)	6 546
Autres (O)	6 747
TOTAL	**26 782**

TABLEAU 2 – Répartition des échantillons par classe dans la base de données d'apprentissage

Nous avons divisé cette base de données en deux parties : 13 films sont utilisés pour constituer notre base d'apprentissage, et 3 sont conservés pour les tests. Pour la construction de la base d'apprentissage, un certain nombre d'échantillons est extrait des films. Ces échantillons sont extraits de manière aléatoire dans chacun des films (les génériques et/ou crédits comprenant des images noires ne sont pas pris en compte), de façon à obtenir au final un nombre d'échantillons équilibré pour chaque classe. Cela nous permet en effet d'éviter que la classe **O** ne domine les trois autres classes. En effet, lorsqu'on déséquilibre le nombre de données des classes dans l'apprentissage, les SVMs ont tendance à favoriser la classe majoritaire. Le tableau 2 contient la répartition des échantillons de la base d'apprentissage par classe. Il est à noter que le nombre d'échantillons total dans la base d'apprentissage reste assez faible.
Pour la base de test, nous avons gardé des films de genres différents :

Armageddon Film contenant principalement des **E** et des **S** ainsi que quelques **G**. Il est plutôt violent.

La mémoire dans la peau Film contenant des combats d'arts martiaux avec beaucoup de chocs et quelques **G**.

Le magicien d'Oz Film sans violence.

Le choix de ces trois films devrait nous permettre de nous faire une idée des performances de l'algorithme sur des vidéos de plusieurs genres, car il est important que l'algorithme fonctionne aussi bien sur des films violents que non

violents. Les tests sont effectués sur tous les échantillons des films, et non plus sur un sous-ensemble.

3.2 Validation croisée

L'utilisation d'algorithmes de classification présuppose dans la plupart des cas l'estimation d'un certain nombre de paramètres. Dans le cas des SVMs il s'agit des paramètres C et γ, représentant le coût appliqué aux erreurs et le paramètre du noyau RBF. Pour estimer ces paramètres, un algorithme de validation croisée et une recherche exhaustive sont utilisés. Cela consiste à séparer en N sous-ensembles les échantillons de la base d'apprentissage, à apprendre sur $N-1$ sous-ensembles, à valider sur le sous-ensemble restant, puis à faire tourner les sous-ensembles de façon à ce que chacun ait pu être testé. Au final, la question latente à cet algorithme est : *Comment organiser les échantillons dans les différents sous-ensembles ?* Les paramètres du modèle sont en général choisis par validation croisée en répartissant les échantillons au hasard, indépendamment de leur provenance. Nous montrons plus loin, de manière empirique, que cette façon de faire n'est pas réaliste. En effet, la validation croisée sera biaisée par le fait que les échantillons des deux bases proviennent des mêmes documents.
Nous faisons différemment dans le sens où nous prenons en compte la provenance des échantillons. Chacun des films de la base d'apprentissage est considéré comme un sous-ensemble, et nous appliquons l'algorithme de validation croisée sur ces sous-ensembles. Cette approche est plus proche des performances du modèle dans la réalité, car chaque test est effectué sur des échantillons dont la provenance est différente de tous ceux qui ont servi à construire le modèle.
Nous nous réfèrerons au fait de répartir les échantillons indépendamment de leur provenance par $CV_{\textit{mélangé}}$ et au fait de considérer les films comme autant de sous-ensembles par $CV_{\textit{séparé}}$. Nous étudions et comparons ces deux approches dans la suite de cet article.

3.3 Méthodes d'évaluation

Pour évaluer les performances des expériences, nous utiliserons différentes mesures. Premièrement, nous pourrons utiliser l'exactitude (accuracy), qui est définie comme suit :

$$A = \frac{1}{K} \sum_{i=1}^{NC} TP_i \qquad (1)$$

où i est l'indice de classe et varie entre $1 \leq i \leq NC$, NC est le nombre de classes, TP_i est le nombre de vrais positifs pour la classe i et K est le nombre d'échantillons considérés. Cette mesure est particulièrement adaptée à la validation croisée [2] dans notre cas, car le nombre d'échantillons de chaque classe dans la base est équilibré. En revanche, pour la base de test, elle l'est beaucoup moins car la classe **O** est fortement majoritaire. L'exactitude sera donc dans ce cas principalement représentative de l'exactitude de la classe **O**.
Pour pallier ce problème, nous utiliserons aussi d'autres mesures comme le rappel et la précision, ou les matrices de confusion.

2. Dans ce cas on parlera d'exactitude validation croisée, ou CVA (Cross-Validation Accuracy).

	S (%)	G (%)	E (%)	O (%)
S	**82,54**	3,87	4,88	8,71
G	3,47	**79,51**	11,87	5,14
E	3,51	7,45	**87,00**	2,03
O	10,29	6,21	3,75	**79,75**

TABLEAU 3 – Matrice de confusion pour la $CV_{\textit{mélangé}}$. La CVA est dans ce cas de 82,17 %.

Rappel et précision relatifs à chaque classe. Le rappel est défini comme étant le rapport entre le nombre de vrais positifs pour la classe i et le nombre d'échantillons de la classe i. La précision est définie comme étant le rapport entre le nombre de vrais positifs pour la classe i et le nombre d'échantillons classés comme appartenant à la classe i.

Matrice de confusion. La matrice de confusion est un outil qui permet de visualiser l'exactitude de chaque classe, ainsi que la répartition des erreurs de chaque classe, appelées confusions entre classes. Les valeurs représentées peuvent être soit des valeurs exactes, soit des proportions.

4 Résultats

Dans cette section, les résultats liés aux deux types de validation croisée sont présentés. Une comparaison est faite entre les deux approches.

4.1 Résultats liés à la $CV_{\textit{mélangé}}$

Les résultats obtenus par la $CV_{\textit{mélangé}}$ sont très bons. Le tableau 3 contient la matrice de confusion du meilleur jeu de paramètres pour cette expérience. La première remarque que l'on peut faire est que l'exactitude des classes est élevée ($\geq$ 79 %). La valeur pour la classe explosions monte même jusqu'à 87 %. Deuxièmement, les confusions entre classes sont relativement faibles, autour de 5-6 %. De plus, on remarque que les confusions les plus fortes sont entre la classe **G** et la classe **E** (ce qui est logique car techniquement un coup de feu est provoqué par une explosion), et entre la classe **S** et la classe **O**.
Ces résultats sont encourageants dans le sens où ils sont obtenus sur une infime portion de signal (40 ms), avec une base de d'apprentissage relativement faible (26 782 échantillons soit $\sim$18 minutes) mais diverse car provenant de 13 films à la fois. Ceci conforte par ailleurs les résultats généralement présentés dans la littérature.
Le tableau 4, quant à lui, contient les résultats des tests effectués sur les trois films de la base de test en terme de précision et rappel pour chacune des classes. On remarque que pour les trois classes minoritaires les résultats sont nettement moins bons. En effet, pour ces trois classes, le rappel

	$CV_{mélangé}$		Armageddon		La mémoire dans la peau		Le magicien d'Oz	
	Précision	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**
S	82,78	82,54	1,19	27,41	0,00	0,00	0 ,08	30,48
G	82,12	79,51	2,13	14,39	2,08	34,95	-	-
E	80,46	87,00	15,6	22,44	0,06	0,79	-	-
O	83,45	79,75	71.05	74,87	68,18	68,47	74,83	74,87

TABLEAU 4 – Rappel et précision (en pourcentages) pour chacune des classes pour la base de test en utilisant le modèle choisi avec la $CV_{mélangé}$.

	$CV_{séparé}$		Armageddon		La mémoire dans la peau		Le magicien d'Oz	
	Précision	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**
S	36,56	35,47	1,51	48,39	0,00	2,86	0 ,15	18,10
G	35,77	33,84	1,89	1,76	2,99	41,52	-	-
E	21,99	8,91	23,24	59,76	0,11	1,59	-	-
O	18,55	73,65	62,96	70,35	68,03	70,33	84,23	84,28

TABLEAU 5 – Rappel et précision (en pourcentages) pour chacune des classes pour la base de test en utilisant le modèle choisi avec la $CV_{séparé}$.

	S (%)	G (%)	E (%)	O (%)
S	**27,41**	5,73	11,47	55,39
G	18,65	**14,39**	12,33	54,63
E	15,76	11,01	**22,44**	50,78
O	18,49	1,67	4,97	**74,87**

TABLEAU 6 – Matrice de confusion d'Armaggedon pour la $CV_{mélangé}$.

	S (%)	G (%)	E (%)	O (%)
S	**35,47**	14,09	4,78	45,66
G	23,12	**33,84**	19,98	23,05
E	25,07	41,54	**8,91**	24,49
O	14,08	6,37	5,90	**73,65**

TABLEAU 7 – Matrice de confusion pour la $CV_{séparé}$. La CVA est dans ce cas de 38,19 %.

n'excède pas 35 % pour chacun de ces films, ce qui signifie que sur les échantillons annotés **S**, **G** ou **E**, le modèle en reconnaît moins de 35 %. De façon à comprendre ces premiers résultats, nous proposons dans le tableau 6 la matrice de confusion pour *Armageddon*. On remarque tout de suite qu'au moins 50 % des échantillons de ces trois classes ont été classés comme appartenant à la classe **O**, qui est la classe majoritaire. On remarque aussi que la confusion entre **G** et **E** ne met plus en valeur le lien physique entre les deux. Enfin, il apparaît aussi que les résultats pour la classe **O** sont plutôt corrects, même si 18 % ont été confondus avec des **S**. Cela correspond à une certaine logique : la classe **O** étant fortement majoritaire, elle est aussi très polymorphe de par sa construction. De plus, les trois évènements violents considérés étant aussi très ciblés, ils sont potentiellement beaucoup plus localisés dans l'espace des descripteurs acoustiques. Ainsi dès que l'on sort des conditions d'apprentissage pour ces trois classes, on a de fortes chances de se trouver dans la classe **O** [3].

Enfin, les valeurs de précision sont extrêmement faibles, voire nulles. Ceci est dû au déséquilibre entre les différentes classes. En effet, même si le rappel de la classe **O** est de 90 %, il y aura toujours des échantillons **O** classés comme **S, G ou E**. Mais même si cela correspond à 2-3 % des échantillons **O**, cela peut correspondre à plusieurs milliers, voire dizaines de milliers d'échantillons... Cela est en général supérieur au nombre d'échantillons dans les autres classes, ce qui explique les très faibles valeurs de précision.

4.2 Résultats liés à la $CV_{séparé}$

Le tableau 7 présente la matrice de confusion pour la meilleure expérience de $CV_{séparé}$, par rapport à l'exactitude. On peut tout de suite remarquer que les valeurs sont semblables à la matrice de confusion d'*Armageddon* dans notre premier protocole expérimental (tableau 6). De fait, le tableau 5 montre que si l'on s'intéresse aux valeurs de précision et de rappel, pour la $CV_{séparé}$, les valeurs sont de l'ordre de celles des films de la base de test. De plus, les résultats pour la base de test sont similaires, voire un peu meilleurs que ceux obtenus avec le jeu de paramètres choisis à l'aide la $CV_{mélangé}$.

On conclut de ces observations que le fait de prendre en compte la provenance des films dans le processus de validation croisée donne un résultat plus proche de la réalité.

3. Pour *Armageddon*, l'exactitude est de 72,21 %, ce qui correspond à peu près au rappel de la classe **O** pour ce film. Au passage, ceci démontre aussi la nécessité de ne pas se limiter à une seule mesure pour analyser un résultat.

5 Discussion

Nous cherchons ici à essayer d'interpréter les résultats de la section 4. En effet, comment expliquer que l'expérience ne donne pas de bons résultats, alors que les résultats de la littérature sur des problèmes similaires sont généralement de l'ordre de 85 % quelle que soit la mesure utilisée ? Pour essayer de faire un diagnostic du problème, nous appliquons notre modèle sur quelques films dont les échantillons ont participé à la construction de la base d'apprentissage. Nous faisons l'hypothèse qu'étant donné le nombre très faible d'échantillons (∼600 par classe) mis dans la base d'apprentissage comparé au nombre d'échantillons total dans le film et à la taille de la base d'apprentissage, leur influence doit être négligeable sur le résultat final.

Les tableaux 8 et 9 contiennent les résultats pour quelques films de la base d'apprentissage en utilisant le modèle de la $CV_{mélangé}$ (resp. le modèle de la $CV_{séparé}$). Il apparaît tout de suite que les résultats sont beaucoup plus satisfaisants en terme de rappel [4]. En regardant les résultats plus en détails, on se rend compte également que les résultats de la $CV_{séparé}$ sont moins bons que ceux de la $CV_{mélangé}$, alors que les résultats sur la base de test nous laissaient plutôt entrevoir le contraire. On peut donc en déduire que la $CV_{séparé}$ a un pouvoir discriminant ou d'adaptation plus important que la $CV_{mélangé}$ qui a un pouvoir de description des données plus important.

La différence de résultats entre les tableaux 4 et 5 et les tableaux 8 et 9 met en évidence un **problème de généralisation**. Nous pensons que les distributions statistiques des évènements que nous essayons de détecter sont rendues différentes d'un film à l'autre par tous les post-traitements audio qui sont mis en œuvre lors de la création du film. La figure 3 contient une illustration du problème dans un hypothétique espace à deux dimensions. Notre intuition est que d'un film à l'autre, en fonction des traitements utilisés, les distributions peuvent devenir tellement différentes qu'il peut n'y avoir plus aucun recouvrement ou que le recouvrement se fasse avec une classe différente. De plus, dans le problème considéré, la situation est encore plus compliquée. Premièrement, nous sommes dans un espace à N dimensions (dans notre cas, $N = 20$). Deuxièmement, les distributions statistiques ne sont pas forcément des clusters de points bien définis et il peut y en avoir plusieurs pour une même classe dans un seul film. Enfin, cette dernière remarque est particulièrement vraie pour la classe **O** car elle contient tout ce qui n'est pas **S**, **G** ou **E**. En bref, cette dernière classe occupe potentiellement tout l'espace statistique entre les trois autres classes.

Pour essayer d'analyser les divergences entre les échantillons des bases d'apprentissage et de test, nous avons calculé la divergence de Jensen-Shannon [10] entre les films et les échantillons de la base d'apprentissage. Cette divergence est une version symétrisée et bornée de la divergence de Kullback-Leibler [5], en modélisant les différentes classes par une loi normale multivariée. Les divergences ont été calculées pour chaque film de la base de test, puis pour chaque film de la base d'apprentissage des tableaux 8 et 9 (pour ceux qui ne possède pas d'échantillons d'une classe, le point n'est pas reporté). La figure 2 met en relation le rappel et la divergence de Jensen-Shannon. On remarque que mis a part quelques points aberrants, il y a une corrélation évidente entre le rappel et la divergence de Jensen-Shannon, ce qui tendrait à confirmer le diagnostic formulé pour notre problème de généralisation.

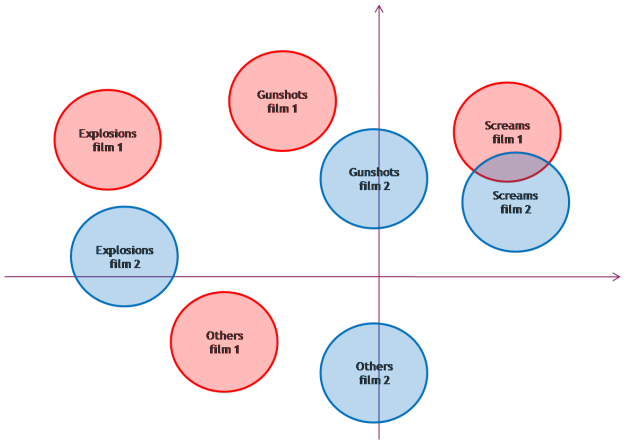

FIGURE 3 – Illustration du problème de généralisation. Les distributions hypothétiques d'un premier film (en bleu) sont comparées à celles d'un deuxième film (en rouge)

De plus, il est intéressant de noter que le faible nombre d'échantillons introduits dans la base d'apprentissage suffit pour obtenir des bons résultats sur les autres échantillons du film. Nous pensons que les SVMs ont réussi à capturer la structure des données statistiques du film avec seulement quelques échantillons. Par ailleurs, ceci contredit notre hypothèse qu'un faible nombre d'échantillons d'un film insérés dans la base d'apprentissage n'aura pas une grande influence sur les résultats d'inférence pour le reste des échantillons du film.

6 Perspectives

Pour essayer de résoudre ce problème de généralisation, plusieurs pistes s'offrent à nous.

Premièrement, l'introduction d'attributs provenant d'une modalité différente, comme la vidéo, permettrait de réunir d'autres informations sur le film pour prendre une décision, et permettrait de vérifier si le problème de généralisation persiste en utilisant la vidéo.

Une autre idée consiste en l'intégration dans le système de la notion de temporalité. Il est en effet logique de penser que les **S**, **G** et **E** ont une structure temporelle particulière qu'il est surement intéressant de prendre en compte pour

4. Les mauvais résultats en terme de précision sont eux dus au déséquilibre entre les trois classes **S**, **G** et **E** et la classe **O**

5. Les valeurs de la divergence sont comprises entre 0 et $\ln 2 \sim 0.693$.

	Il faut sauver le soldat Ryan		Je suis Légende		Harry Potter 4		Reservoir Dogs	
	Précision	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**
S	3,14	82,51	25,69	78,54	1,45	97,63	4,67	98,10
G	14,20	62,02	12,22	95,73	-	-	18,99	99,37
E	7,05	95,22	16,76	97,94	0,92	98,15	-	-
O	45,35	48,11	79,98	84,32	83,36	83,53	86,06	87,06

TABLEAU 8 – Rappel et précision (en pourcentages) pour chacune des classes pour certains films de la base d'apprentissage en utilisant le modèle choisi avec la $CV_{mélangé}$.

	Il faut sauver le soldat Ryan		Je suis Légende		Harry Potter 4		Reservoir Dogs	
	Précision	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**	**Précision**	**Rappel**
S	2,34	62,10	22,82	68,14	1,33	87,24	3,92	83,89
G	9,09	29,08	13,01	80,53	-	-	24,48	92,13
E	4,79	91,14	9,26	75,75	0,6	100	-	-
O	42,77	45,36	79,93	84,27	81,72	81,89	86,78	87,78

TABLEAU 9 – Rappel et précision (en pourcentages) pour chacune des classes pour certains films de la base d'apprentissage en utilisant le modèle choisi avec la $CV_{séparé}$.

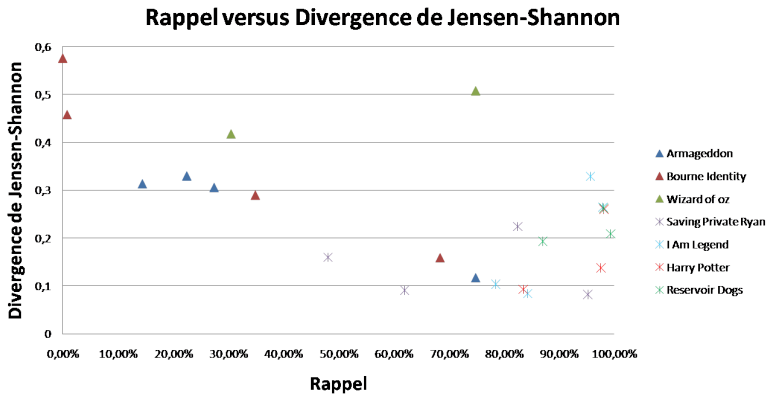

FIGURE 2 – Rappel versus Divergence de Jensen-Shannon par film pour chacune des classes présentes dans le film. Les croix correspondent aux films ayant servi à la construction de la base d'apprentissage, tandis que les triangles correspondent aux films de la base de test.

prendre notre décision. L'introduction de temporalité peut se faire à travers l'utilisation de techniques de classification qui la prennent en compte telles que les modèles de Markov, ou les réseaux bayésiens. Dans [8], les auteurs ont fait une étude de différents cas d'introduction de la temporalité, à travers l'intégration précoce, constituée par l'agrégation des attributs au travers de statistiques, ou l'introduction tardive, constituée par les techniques de classification, et ont montré l'utilité de l'information contenue dans l'évolution des caractéristiques du signal.

Enfin, une autre idée intéressante serait d'utiliser une technique d'adaptation de domaine [4]. Le problème est le suivant : on a un premier jeu de documents d'une provenance X sur lequel on a réussi à faire un premier modèle. On récupère alors un nouveau jeu de documents similaires mais d'une provenance différente Y. Le principe est d'essayer d'utiliser les informations contenues dans X pour essayer de trouver celles contenues dans Y sans avoir à annoter Y (ou le moins possible). Le problème est que les distributions statistiques des évènements dans X sont liées mais souvent différentes de celles dans Y. On utilise une technique d'adaptation de domaine pour essayer de résoudre ce problème. Dans notre cas, l'idée serait de considérer chacun des films comme autant de domaines différents et d'appliquer un algorithme d'adaptation pour essayer de trouver les **S**, les **G** et les **E** dans les films de la base de test.

7 Conclusion

Dans cet article, nous avons étudié la question de la classification d'évènements audio violents dans des films. Nous avons utilisé un système équivalent à l'état de l'art, et

choisi les SVMs comme technique de classification. Nous avons montré que notre système avait de faibles performances sur la base de test. Pour cela, nous avons étudié deux axes. Premièrement, nous avons montré que la $CV_{mélangé}$ n'est pas adaptée à un problème comprenant des échantillons audio provenant de plusieurs films. En effet, la $CV_{mélangé}$ nous donne une CVA de 82 %, tandis que les tests montrent qu'elle est plutôt de l'ordre de 20-30 % en réalité. En considérant nos différents films comme autant de sous-ensembles et en les utilisant pour faire notre validation croisée, nous avons montré que le résultat est dans ce cas beaucoup plus proche de la réalité.
Deuxièmement, nous avons mis en lumière un problème de généralisation, dû au fait que les différentes distributions statistiques des évènements choisis recherchés dans les films ne sont pas identiques. Ce problème est lié à ce que nous avons précédemment montré car la $CV_{séparé}$ nous donne une meilleure estimation du pouvoir discriminant du modèle et de ses performances en généralisation. Nous avons également confirmé que ce problème est dû à une disparité des distributions statistiques des classes en calculant la divergence entre la base d'apprentissage et les films de la base de test et montré une corrélation avec le taux de rappel.

Références

[1] Miguel Bugalho, José Portelo, Isabel Trancoso, T Pellegrini, and A Abad. Detecting audio events for semantic video search. In *InterSpeech*, 2009.

[2] Liang-Hua Chen, Chih-Wen Su, Chi-Feng Weng, and Hong-Yuan Mark Liao. Action scene detection with support vector machines. *Journal of Multimedia*, 4 :248–253, 2009.

[3] Ankur Datta, Mubarak Shah, and Niels Da Vitoria Lobo. Person-on-person violence detection in video data. In *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1 - Volume 1*, ICPR '02, pages 10433–, Washington, DC, USA, 2002. IEEE Computer Society.

[4] Hal Daumé, III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1) :101–126, 2006.

[5] Theodoros Giannakopoulos, Dimitrios I. Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. Violence content classification using audio features. In *4th Helenic Conference on Artificial Intelligence*, pages 502–507, 2006.

[6] Theodoros Giannakopoulos, Dimitrios I. Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. A multi-class audio classification method with respect to violent content in movies using bayesian networks. In *IEEE 9th Workshop on Multimedia Signal Processing*, pages 90 –93, oct. 2007.

[7] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In Stasinos Konstantopoulos, Stavros Perantonis, Vangelis Karkaletsis, Constantine Spyropoulos, and George Vouros, editors, *Artificial Intelligence : Theories, Models and Applications*, volume 6040 of *Lecture Notes in Computer Science*, pages 91–100. Springer Berlin / Heidelberg, 2010.

[8] Cyril Joder, Slim Essid, and Gael Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1) :174 –186, jan. 2009.

[9] Bart Lehane, Noel E. O'Connor, and Noel Murphy. Action sequence detection in motion pictures. In *European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*, 2004.

[10] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37 :145–151, 1991.

[11] Simon Moncrieff, Chitra Dorai, and Svetha Venkatesh. Affect computing in film through sound energy dynamics. In *ACM Multimedia*, pages 525–527, 2001.

[12] Simon Moncrieff, Chitra Dorai, and Svetha Venkatesh. Detecting indexical signs in film audio for scene interpretation. In *Proceedings of the International Conference on Multimedia & Expo*, pages 989 – 992, aug. 2001.

[13] Simon Moncrieff, Svetha Venkatesh, and Chitra Dorai. Horror film genre typing and scene labeling via audio analysis. In *Proceedings of the International Conference on Multimedia & Expo*, volume 2, pages 193–196, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[14] Aggelos Pikrakis, Theodoros Giannakopoulos, and Sergios Theodoridis. Gunshot detection in audio streams from movies by means of dynamic programming and bayesian networks. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 21–24, 2008.

[15] Ba Tu Truong and Svetha Venkatesh. Determining dramatic intensification via flashing lights in movies. In *Proceedings of the International Conference on Multimedia & Expo*, pages 60 – 63, aug. 2001.

[16] Shuhui Wang, Shuqiang Jiang, Qingming Huang, and Wen Gao. Shot classification for action movies based on motion characteristics. In *Proceedings of the International Conference on Image Processing*, pages 2508 –2511, oct. 2008.

[17] Lexing Xie, Hari Sundaram, and Murray Campbell. Event mining in multimedia streams. *Proceedings of the IEEE*, 96(4) :623 –647, april 2008.