Contributions à la détection de concepts et d'événements dans les documents vidéos

Nadia Derbas

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministériel : 7 août 2006

Présentée par

Nadia Derbas

Thèse dirigée par **Georges Quénot**

préparée au sein du **Laboratoire d'Informatique de Grenoble** dans **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII)**

Contributions à la détection de concepts et d'événements dans les documents vidéos

Thèse soutenue publiquement le **« 30 Septembre 2014 »**, devant le jury composé de :

M. Guillaume Gravier
Chargé de recherche CNRS, CNRS, Rapporteur
M. Philippe Joly
Professeur, Institut de Recherche en Informatique de Toulouse, Rapporteur
Mme. Claire-Hélène Demarty
Chercheuse, Technicolor, Examinateur
M. Patrick Lambert
Professeur, Université de Savoie, Examinateur
M. Georges Quénot
Chargé de recherche CNRS, CNRS, Directeur de thèse

Remerciements

Ces quatre années de thèse ont été pour moi une aventure très riche d'un point de vue humain et scientifique. Il est particulièrement difficile de remercier toutes les personnes qui m'ont aidé à mener à bien mon projet, pour cela je voudrais commencer par remercier toutes les personnes que j'ai croisé tout au long de ma thèse.

Tout d'abord, je remercie mon directeur de thèse Georges Quénot, qui m'a donné l'occasion de vivre cette aventure. Je te remercie, Georges, pour ton suivi , pour tes grandes connaissances scientifiques et pour la liberté de travail que tu m'a laissé et qui m'a permis de développer mon autonomie. Je remercie le projet OSEO-Quaero pour le financement fourni qui m'a permis d'accomplir ce travail.

Je remercie ensuite les membres du jury, Patrick Lambert, Philippe Joly, Claire-Hélène Demarty et Guillaume Gravier d'avoir accepté d'évaluer mon travail et pour leurs commentaires.

Je remercie également tous les membres de l'équipe MRIM qui m'a accueilli pendant ma thèse ainsi que Franck, Émilie et Martine, pour leur bonne humeur et les bons moments partagés ensemble, réunions, conférences, repas de noël, ... Et je remercie plus particulièrement Catherine pour tous les échanges que nous avons pu avoir tout au long de ma thèse et pour ses conseils avisés.

Je tiens à remercier Laurent Besacier pour son soutien et ses encouragements dans les moments difficiles de la thèse.

Je voudrais remercier mes collègues du LIG ou plutôt mes « frères d'armes » qui m'ont accompagné tout au long de ma thèse et qui ont égayé mes journées : Sarah, Quang, Pathathai, Uyanga, Mateusz, Bahjat, Rami, Bachar, Abdelkader, Vanildo, Fred, David... Nos repas de midi, pauses cafés et soirées ont été des vrais moments de partage, de plaisir et de découverte de nouvelles cultures.

Je remercie bien évidement mes amies les plus proches de leur présence : Amélia, Elisa, Dominika, Haitang et Lorène, qui ont toujours su m'écouter, me distraire et m'encourager dans ce que j'entreprends.

« The last but not the least », je remercie ma chère famille. Maman, Papa, Dima, Darine, Mohamad et Bisso, vous m'avez toujours soutenu et cru en moi. Vous avez été à mes côtés dans les moments les plus heureux mais aussi les plus difficiles. Vous m'avez appris à être ambitieuse et rigoureuse, c'est grâce à vous que j'ai pu aller aussi loin.
Enfin, je veux adresser les remerciements qui me tiennent le plus à cœur, ceux pour mon ami, mon amour, mon mari, mon tout, Yann pour son soutien et son amour inconditionnel. Tu as été le premier à me réconforter et à me pousser au quotidien. Sans toi, cette aventure n'aurait pas été la même.

Table des matières

1 Introduction

Les moyens de communications ont connu un bouleversement radical avec la démocratisation de la technologie numérique durant cette dernière décennie. En effet, la baisse des prix des appareils photos numériques et leur intégration quasi-systématique aux téléphones mobiles a rendu les appareils photos omniprésents dans la vie quotidienne des consommateurs. Les consommateurs ont alors acquis de nouveaux réflexes en capturant des images ou des vidéos quand ils veulent, partout où ils vont et en capturant des scènes auxquelles ils ne se seraient jamais intéressés avant la démocratisation des téléphones mobiles. Aujourd'hui, il y a plus de photos prises par des téléphones mobiles que par des appareils photos. En effet, sur les 5,2 milliards de mobiles actuellement en service, 4,4 milliards sont dotés d'un appareil photo.

L'émergence des plateformes de partage en ligne et de réseaux sociaux (Youtube, Dailymotion, Flickr, Snapchat, Instagram, Vine, Facebook, ...) a modifié les comportements sociaux des utilisateurs. Par conséquent, les utilisateurs créent et partagent de plus en plus de documents multimédias. La plateforme de partage de vidéos en ligne créée en 2007, Youtube, a annoncé plus de 100 heures de vidéos téléversées chaque minute sur leur plateforme en 2013 contre 35 heures de vidéos téléversées chaque minute en 2010[1]. La plateforme de partage d'images en ligne Flickr a annoncé que le nombre d'images téléversées par jour s'est multiplié par trois en 2013 pour atteindre plus de 1,6 millions d'images téléversées par jour[2]. La tendance est la même chez Facebook où plus de 350 millions de photos sont téléversées chaque jour. Ceci fait de Facebook la plus grande « bibliothèque de photos au monde », avec 250 milliards d'images, soit près de 30 fois plus que Flickr et 70 fois plus qu'Instagram. Pour avoir un ordre d'idée, le nombre total de photos mises en lignes et partagées

[1] http://www.youtube.com/yt/press/statistics.html
[2] http://blog.flickr.net/en/2014/02/06/a-note-of-thanks-2/

quotidiennement en 2013 a été évalué à 1,2 milliards selon le rapport annuel sur les tendances chiffrées d'Internet de Mary Meeker [Meek 13].

Par ailleurs, la décroissance très rapide des coûts de stockages et la hausse constante de l'utilisation du Cloud, ont incité les utilisateurs à troquer l'archivage papier ou assimilé pour l'archivage numérique. En effet, les coûts de stockage sont en baisse en moyenne de 38% par an depuis 1992. À titre d'exemple, les disques durs externes affichent, de nos jours, des prix très abordables pour de très grandes capacité. De plus, le recours aux nouveaux services de stockage et d'archivage en ligne (Cloud) explosent (comme ceux proposés par Amazon, Dropbox), par exemple le nombre de documents stockés sur Amazon S3 est passé de 0 à 2 milliards en 5 ans.

Les documents multimédias cités précédemment sont probablement les plus volumineux par rapport à d'autres types de documents qui ont également vu leur nombre exploser et qui sont bien plus importants en termes d'intérêt. En effet, l'explosion de la quantité des documents multimédias touche différents domaines : l'entreprise, le journalisme, l'architecture, la publicité, la médecine, le cinéma, la télévision ... L'Institut National de l'Audiovisuel (INA) a quantifié le flux télévisuel archivé tous les ans à 930 000 heures. Selon l'enquête internationale sur les statistiques des films de long métrages, menée par l'Institut de Statistique de l'UNESCO (ISU) et publiée en 2013, la production mondiale annuelle de films s'est intensifiée, augmentant ainsi de 39% entre 2005 et 2011 et passant de 4 818 à 6 573 longs métrages.

Enfin, la popularisation de la technologie numérique, les nouveaux comportements sociaux, la multiplication des plateformes de partage et le développement rapide de l'accès en temps réel aux données ont contribué à l'expansion de la création de données et à la très forte croissance des volumes des documents multimédias disponibles.

1.1 Besoin d'automatisation

Des méthodes ont été mises en place pour faciliter la recherche et la manipulation des documents multimédias. Ces méthodes consistent à associer aux collections de documents multimédias un index décrivant leur contenu. Ceci est une étape cruciale pour les applications de gestion ou de recherche de données multimédias qui jusqu'à maintenant se fait, le plus souvent, manuellement par les producteurs des documents ou par des documentalistes qui attribuent un certain nombre d'étiquettes aux documents multimédias.

L'explosion du volume d'images et de vidéos a rendu cette indexation des documents multimédias très coûteuse et manuellement impossible. Par conséquent, il apparaît nécessaire de disposer de systèmes capables d'analyser, de stocker et de retrouver les documents multimédias automatiquement en leur attribuant des étiquettes et donc d'avoir un système d'indexation automatique de contenus multimédias. Un tel

système analyse des documents multimédias et détermine de leur contenu sémantique (figure 1.1).

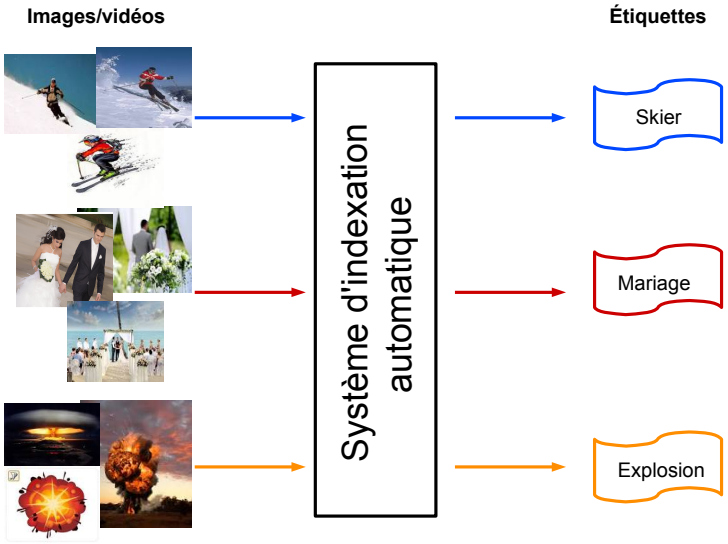


FIGURE 1.1 – Un système d'indexation automatique

L'indexation automatique est très utile pour de nombreux domaines d'applications. Par exemple, pour l'organisation des vidéos et des images des plateformes de partage (Youtube, Flickr, Instagram, ...) et des collections personnelles et pour faciliter donc, la recherche de documents, sans que les propriétaires n'aient à ajouter manuellement des étiquettes. Elle peut également être utile pour l'archivage du flux télévisuel en facilitant la recherche de vidéos anciennes reliées à un nouveau fait d'actualité, en permettant la reconnaissance de genres des vidéos et des films (drame, comédie, documentaires, ...) et en segmentant les vidéos de sport comme les matchs de football de façon à trouver les séquences comportant le plus d'action intéressantes (par exemple les séquences comportant les buts). L'indexation est également très utile pour la lutte contre le piratage et protéger les droits d'auteurs de documents multimédias (par exemple de musique ou de films, ...). Mais elle l'est aussi pour les systèmes de vidéo-surveillance en détectant des événements spéciaux ou suspects.

Ces exemples montrent l'importance de l'indexation automatique de contenus multimédias. En revanche, toutes les techniques d'indexation actuelles rencontrent des problèmes de faisabilité et leur qualité est variable selon les applications. Par exemple, il est impossible de différentier certains objets complexes, comme identifier la provenance de pollens en se basant sur une photographie microscopique. Ces systèmes ont également un taux de précision relativement faible. À titre d'exemple, effectuer une recherche de vidéos sur l'animal « chat » pourra, avec certains systèmes, retourner un grand nombre de résultats avec une faible proportion de vidéos contenant des chats (bon rappel, faible précision). À l'opposé, d'autres systèmes vont re-

tourner un petit nombre de résultats avec une grande proportion de vidéos contenant des chats (bonne précision, faible rappel).
L'objectif principal des travaux de recherche sur l'indexation automatique est d'élargir son champ d'application et d'améliorer ses performances pour pouvoir satisfaire plus d'applications ou de rendre plus satisfaisantes les applications existantes. L'amélioration de la performance de ces systèmes permettra une baisse des coûts de l'extraction de l'information et entraînera l'apparition d'une nouvelle génération d'applications d'analyse du contenu. Dans cette thèse, nous nous intéressons à l'amélioration de la performance globale de ces systèmes d'indexation.

1.2 Système d'indexation par le contenu

Un système d'indexation attribue automatiquement une ou plusieurs étiquettes à des échantillons ou des unités d'indexation et de recherche multimédia comme des images fixes, des vidéos entières, des plans vidéos ou des segments vidéos correspondant à des unités sémantiques (par exemple reportage ou sujet dans les journaux télévisés. Il l'effectue en analysant le contenu de ces derniers sans aucune intervention humaine. Il détecte des concepts pour enrichir les documents multimédias (image ou vidéo) par une description textuelle qui résume leur contenu sémantique. Plus formellement, soit $X = x_1, ..., x_n$ un ensemble d'échantillons (ou instances) et $C = c_1, ..., c_k$ un ensemble de concepts. L'objectif est d'apprendre une fonction de prédiction $f : X \rightarrow [0, 1]^k$ qui prédit la probabilité de présence de chaque concept dans un échantillon. Ces concepts peuvent être statiques (comme des objets) ou dynamiques (comme des mouvements, actions ou événements). Il n'existe pas de définitions précises pour les différents types de concepts.

Dans la littérature, il existe un grand nombre de travaux concernant les systèmes d'indexation dont les supervisés sont les plus communs. Les systèmes d'indexations supervisés entraînent des modèles par apprentissage supervisé sur un ensemble d'échantillons annotés manuellement pour chaque concept considéré (les données d'apprentissage). La figure 5.2 montre le processus d'un système d'indexation supervisé classique et ses différentes étapes. Ce processus est composé principalement de deux grandes étapes : l'apprentissage des classificateurs et la prédiction. L'apprentissage des classificateurs consiste à extraire des descripteurs qui transforment l'information contenue dans les échantillons en une représentation plus adaptée à l'apprentissage (par exemple des descripteurs couleurs, texture, audio ...) ; et d'apprendre ensuite une fonction de prédiction (ou modèle d'apprentissage) à partir des descripteurs de l'ensemble d'apprentissage et des étiquettes qui leur sont attribuées. La prédiction, quant à elle, consiste à utiliser la fonction de prédiction apprise sur les données d'apprentissage pour attribuer un score (probabilité de présence d'un concept) à un nouvel échantillon en fournissant ses descripteurs. Ce processus peut être appliqué plusieurs fois sur les mêmes données mais avec différentes variantes, ainsi des méthodes de fusion existent pour intégrer les informations apportées par les différents systèmes.

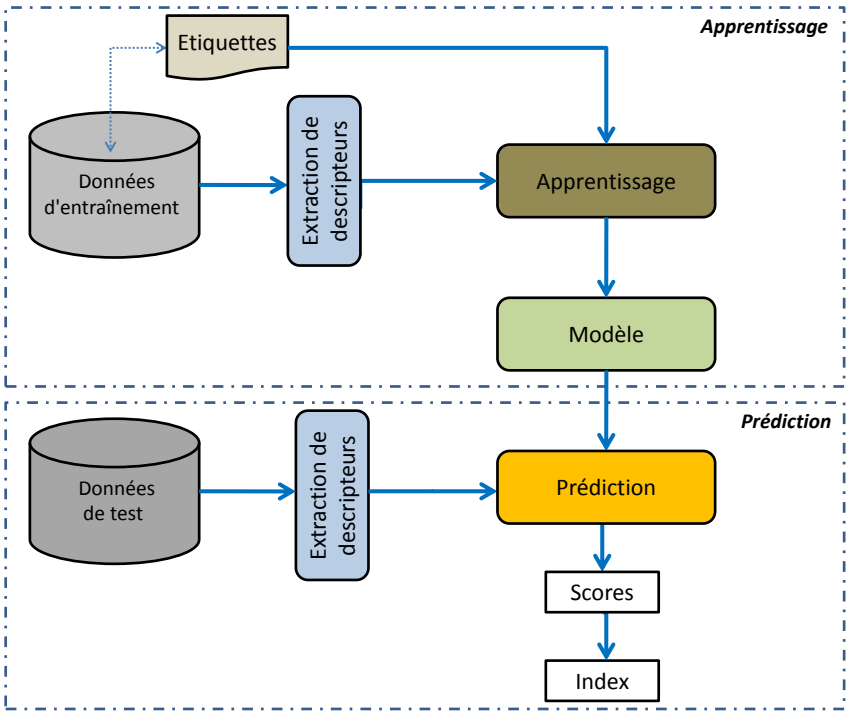


FIGURE 1.2 – Le processus général d'un système d'indexation classique.

La performance des systèmes d'indexation reste très limitée et les systèmes disponibles de recherche et de gestion des documents multimédias (comme les moteurs de recherche) dépendent encore d'une attribution d'index et d'étiquettes manuelle. Les systèmes d'indexation automatiques doivent surmonter de nombreuses difficultés pour pouvoir intégrer les produits commerciaux. Parmi ces difficultés, il y a la variabilité dans les données à traiter : pour gagner de la fiabilité et de l'efficacité, le système doit être capable de gérer la variabilité des environnements, de formes, de positions, de poses, d'illuminations, d'orientations (comme le montre la figure 1.3). Les grands volumes de données constituent quant à eux la deuxième difficulté : pour supporter toute la masse de documents multimédias produites, le système doit être capable de passer à l'échelle et traiter de très grands volumes de données tout en respectant des contraintes de temps de calcul et de stockage.

FIGURE 1.3 – La variabilité de forme, d'environnement et de positions pour le concept vélo.

1.3 Principales contributions de cette thèse

La performance des systèmes d'indexation automatique de contenus multimédias dépend de plusieurs facteurs. Parmi les facteurs les plus influents nous citons : la qualité des données d'apprentissage, la capacité des descripteurs choisis pour représenter le contenu des documents, la capacité de la méthode de classification utilisée à apprendre les corrélations entre les classes et les représentations de chacune d'elles et enfin la faculté des méthodes de fusion des différentes représentations à prendre en compte la corrélations entre les différents modalités.

Ce travail de thèse s'est déroulé dans le cadre du projet *QUAERO* et ses sous-tâches sur la structuration multimodale et la reconnaissance d'objets et d'événements dans les vidéos.

Dans cette thèse, nous avons orienté notre travail sur l'amélioration générale de la performance des systèmes d'indexation automatique des documents multimédias par leur contenu. Pour cela nous nous avons abordé ce problème sous différents angles :

- Dans le cadre de la fusion de différentes modalités ou sources d'information : un des problèmes des méthodes de fusion de l'état de l'art est la perte potentielle de corrélation entre les différentes modalités. Par conséquent, nous proposons une méthode de fusion pour représenter conjointement le contenu audio-visuel dans le contexte de la détection automatique de scènes violentes. Cette méthode découvre des motifs audio-visuels spécifiques en construisant un dictionnaire audio-visuel joint (Chapitre 3).

- Dans le cadre de la localisation des concepts dans les images : nous avons voulu utiliser le moins d'annotations manuelles possibles pour réduire les coûts liés aux annotations manuelles précises des vidéos. Nous proposons donc une méthode de localisation faiblement supervisée. Cette méthode consiste à détecter de l'invariabilité spécifique à un concept donné dans la variabilité globale d'une vidéo dans des données d'apprentissage faiblement annotées. Elle permet, ainsi, de localiser les concepts spatialement et temporellement dans les vidéos (Chapitre 4).

- Dans le cadre de la réduction du bruit généré par des annotations ambiguës des données d'apprentissage : deux nouvelles méthodes ont été proposées pour réduire ce bruit. Elles utilisent le contenu visuel des plans et des vidéos et se basent sur l'idée que « Les plans contenant un concept ou un événement sont semblables alors que les plans ne représentant pas ce concept ou cet événement sont différents entre eux et du reste des plans ». Nous présentons une première méthode qui produit des nouvelles annotations et une deuxième qui, elle, pondère les annotations en fonction de la confiance attribuée à cha-

cune d'elles (Chapitre 5).

- Dans le cadre de l'optimisation des représentations du contenu multimédia : Une très grande variété de descripteurs existent. Cependant, les descripteurs les plus efficaces possèdent souvent des caractéristiques (comme leur grande dimension) qui les rendent difficilement utilisables sur les grande collections de données. Nous proposons une méthode d'optimisation de descripteurs dédiés aux systèmes d'indexation et de recherche de documents multimédias par le contenu. La méthode proposée combine différentes transformations pour trouver un compromis entre la dimension des descripteurs et leur capacité à représenter le contenu (Chapitre 6).

L'évaluation de nos différentes contributions forme une partie prépondérante de ce travail. Nous avons voulu proposer des méthodes capables de traiter de très grands corpus (Big Data) dans des temps raisonnables. La participation aux campagnes d'évaluation (TRECVid et MediaEval) et l'implémentation optimisée ont donc été au cœur de cette thèse.

Le reste de la thèse est organisée comme suit : le Chapitre 2 présente l'état de l'art concernant le processus d'indexation et les domaines de contributions. Les Chapitres 3, 4, 5 et 6 décrivent nos quatre principales contributions. Enfin, le Chapitre 7 conclut la thèse et expose les perspectives possibles.

2
État de l'art

Dans ce chapitre, nous proposons un tour d'horizon du domaine d'indexation automatique du contenu des documents multimédias. Nous commençons par présenter les données que nous traiterons le long de cette thèse : les documents multimédias. Ensuite nous décrivons la phase de segmentation intervenant en amont de tout processus d'indexation et qui prépare les documents multimédias à l'indexation. Nous développons ensuite le système d'indexation que nous avons adopté le long de cette thèse. Et nous proposons une étude globale des différentes méthodes de l'état de l'art pour chacune des étapes d'un processus classique d'indexation automatique du contenu des documents multimédias. Nous abordons également l'indexation par apprentissage profond ou « Deep Learning » qui trouve de plus en plus de succès dans le domaine de l'indexation de documents multimédias par le contenu.
Enfin, nous présentons un panorama comparatif des collections de données et des campagnes d'évaluation les plus populaires dans le domaine de l'indexation automatique des documents multimédias.

2.1 Documents multimédias

Le terme « documents multimédias » est apparu vers la fin des années 1980 pour désigner les documents mettant en œuvre l'image fixe ou animée, le son, le texte. L'image fixe est définie par une représentation visuelle de quelque chose (objet, être vivant et/ou concept). Une des plus anciennes définitions de l'image est celle de Platon : « j'appelle image d'abord les ombres ensuite les reflets qu'on voit dans les eaux, ou à la surface des corps opaques, polis et brillants et toutes les représentations de ce genre ». La vidéo, quant à elle, est définie comme étant une combinaison de différents types de flux, essentiellement d'un flux visuel et d'un flux audio (son). Le flux visuel est formé d'enchaînement de séquences d'images fixes à raison de 25 à 60 images par secondes. Le flux sonore peut comporter un seul canal (mono) ou plusieurs canaux (stéréo ou des systèmes de généralisation de la stéréo comme 5.1 ou 7.1). Pour numériser le flux sonore, le signal analogique est échantillonné entre 11 000 Hertz et 48 000 Hertz. La fréquence la plus souvent utilisée est de 44 100 Hertz. Enfin, un troisième flux d'information qui peut être associé aux vidéos est le texte. Ce flux textuel est présent dans certaines vidéos pour faciliter la compréhension du contenu ou pour apporter des précisions supplémentaires concernant la vidéo comme par exemple le nom de l'auteur, la date, le lieu ...

La vidéo est le document multimédia le plus complexe à traiter à cause de la quantité d'information visuelle, audio et textuelle qu'elle peut contenir, ainsi que sa dimension temporelle et spatiale. Ceci rend son volume bien plus grand que celui d'une image fixe ou d'un document texte. Au vu de la complexité des vidéos, une étape de segmentation s'est imposé pour permettre une meilleure analyse de son contenu.

2.2 Segmentation de vidéos

Les vidéos sont des flux de données non structurés et constitués d'une succession d'images de longueur variable pouvant aller de quelques secondes à quelques heures. Cette longueur et richesse de contenu rend un accès efficace aux vidéos et leur traitement très difficile. Une des approche les plus répandues pour la représentation du contenu des vidéos est la modélisation structurée. Cette approche consiste à découper les vidéos en plus petites séquences (dites plans). Un plan est une séquence d'images générées par une prise de caméra sans coupure ou interruption et présentant une action continue.

En 1991, Thomas *et al.* proposent un modèle de représentation du contenu des vidéos pour permettre aux réalisateurs d'accéder rapidement aux différentes séquences de films et de voir leurs descriptions [Smit 93]. Ce modèle est baptisé « stratification » ou modèle en couches. La caractéristique de ce modèle est qu'il ne représente pas la vidéo en tant qu'ensemble de plans indépendants mais plutôt comme un ensemble de strates entrelacées. Une strate est une information descriptive du contenu, du contexte, du son ou de la perspective en quelques mots, par exemple : « John descend la colline » ou encore « fleurs dans le vent ». Les strates sont organisées hiérarchiquement : le premier niveau contient toutes les images de la vidéo, puis plus on monte dans la hiérarchie plus les séquences sont regroupées entre elles selon leur contenu sémantique [Dave 91]. Ce modèle a été largement repris par la suite pour structurer les documents vidéos [Chua 02, Kank 00, Rui 98, Weis 95].

La figure 2.1 illustre cette structure hiérarchique en strates d'un document vidéo. la première strate contient toutes les images de la vidéo. Ensuite les images sont regroupées en plans avec une possibilité d'extraire pour chacun des plans une image clé représentant le plan. Les plans partageant des attributs communs sont regroupés sous forme de scènes. Enfin les scènes reliés sémantiquement sont regroupées sous forme d'histoires.

L'indexation de texte nécessite l'utilisation de mots et de phrases comme pointeurs sur des paragraphes, des pages ou des documents entiers. De la même façon, l'indexation de vidéo nécessite la sélection d'images clés et de plans comme pointeurs pour des unités de niveaux supérieurs comme des scènes et des histoires. L'unité la plus basique, après l'image fixe, des systèmes d'analyse par le contenu des documents vidéos est le plan [Naph 04, Over 05]. Ainsi, une segmentation fiable et précise des vidéos en plans est devenue une étape clé dans l'indexation automatique des vidéos. Une fois que les limites des différents plans d'une même vidéo sont détectées, la seule image qui sera gardée pour un plan donné est l'image la plus représentative du plan, dite l'image clé. Ceci permettra de réduire la quantité d'information à traiter par les systèmes d'indexation automatiques par la suite, même si certaines approches préfèrent utiliser plusieurs images-clés par plan.

La segmentation en plans a une importance fondamentale pour de nombreuses ap-

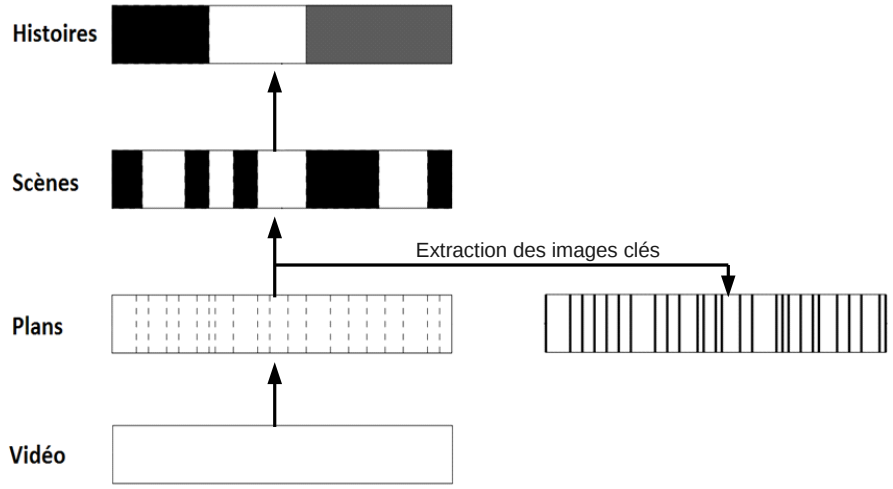


FIGURE 2.1 – La structure hiérarchique en strates d'un document vidéo.

plications d'analyse, d'indexation et de recherche par le contenu dans une collection de documents vidéos. Les changements de plans sont reconnaissables par un changement brutal d'images ou encore par des transitions de dégradé ou d'estompage. Cette segmentation s'appuie généralement sur une analyse de l'information visuelle de la vidéo [Bore 96]. Au vu de l'importance de la segmentation, une tâche à part entière lui a été consacrée pour les campagnes d'évaluation TRECVid de 2001 à 2007. Le but de cette tâche est de proposer des systèmes capables de détecter automatiquement les limites des plans dans les documents vidéos. Smeaton *et al.* ont retracé les différents systèmes proposés tout au long de ces sept années d'évaluation en les comparant et en détaillant ceux qui ont été les plus performants [Smea 10]. Parmi ces systèmes on trouve notamment, le système proposé par Quénot qui est basé sur la détection de deux types de transition : la coupure nette et le fondu enchaîné [Quen 01]. Ce système a fourni la segmentation officielle des vidéos des différentes tâches de TRECVid durant plusieurs années.

Durant ces dernières années, la segmentation en plans a été largement explorée. Un grand nombre de systèmes très précis ont alors été proposés et adoptés comme la base des systèmes d'indexation automatique des documents vidéos. Actuellement, la recherche se tourne vers une segmentation plus complexe, en segments plus signifiants et contenant plus de sémantique, qui est la segmentation en histoires [Dumo 12, Lu 10, Feng 12, Khou 14].

2.3 Processus du système d'indexation des vidéos

Les systèmes d'indexation attribuent automatiquement une ou plusieurs étiquettes aux vidéos en analysant leur contenus. Les étiquettes correspondent aux concepts retrouvés dans les vidéos. Ces systèmes disposent souvent de toute une batterie de concepts à retrouver en même temps dans les vidéos. Par conséquent, certaines étapes du processus d'indexation sont communes à tous ces concepts alors que le reste du processus est souvent adapté à chacun des concepts. Dans la littérature, il existe un grand nombre de travaux concernant les systèmes d'indexation et la détection automatique d'objets ou d'événements. Parmi ces systèmes, nous pouvons citer les systèmes classiques basés sur l'apprentissage en surface et ceux qui sont basés sur l'apprentissage en profondeur (Deep Learning).

Les systèmes basés sur l'apprentissage en surface sont considérés comme les systèmes d'indexation classiques. Ces systèmes reposent sur la répétition sans la compréhension. Ils ont tous la même base quel que soit le type de données (images ou vidéos) et quel que soit le type de concepts considérés (objets ou événements). Cette base est constituée de deux grandes parties : la classification qui extrait des descripteurs représentant le contenu des documents multimédias puis apprend une fonction de prédiction (ou un modèle d'apprentissage) à partir des descripteurs extraits ; et la prédiction qui à l'aide de la fonction apprise attribue des étiquettes à un nouvel échantillon.
La classification supervisée est la plus commune, elle entraîne des modèles par apprentissage supervisé sur les données d'apprentissage (*i.e.* sur un ensemble d'échantillons annotés). Les échantillons de l'ensemble d'apprentissage sont manuellement annotées pour chaque concept considéré.
Les systèmes fondés sur l'apprentissage en profondeur, quant à eux, reposent sur le fonctionnement du cerveau humain. Ils proviennent de la recherche sur les réseaux de neurones artificiels et permettent de modéliser les représentations. Ils s'inspirent du système visuel humain qui est hiérarchique par nature, ces systèmes effectuent donc des transformations non-linéaires en plusieurs couches pour apprendre les différents niveaux des représentations. Nous exposons plus en détails ces systèmes dans la section 2.8.

Dans cette thèse, nous nous plaçons dans le cadre d'un système d'indexation classique fondé sur un apprentissage en surface supervisé. Le processus d'indexation que nous avons adopté dans nos travaux et que nous avons utilisé pour les différentes expérimentations possède quelques étapes supplémentaires. C'est un système en six étapes que nous illustrons dans la figure 2.2 et que nous détaillons ci-dessous :

1. ***Extraction de descripteurs :*** les descripteurs sont calculés pour représenter les différentes informations contenues dans une image (ou vidéo) : des caractéristiques visuelles, des sons spécifiques et du mouvement ou les trajectoires dans les vidéos. Plus de détails concernant ces descripteurs sont donnés dans la section 2.4.

2. ***Optimisation de descripteurs :*** L'optimisation des descripteurs permet de former des descripteurs plus compactes et plus précis pour la représentation du contenu. Ce post-traitement des descripteurs permet à la fois d'améliorer leur performance et de réduire leur taille. L'optimisation des descripteurs peut se faire de différentes façons : par une normalisation ou par une réduction de dimensions des descripteurs. Plus de détails concernant les méthodes d'optimisation sont donnés dans la section 2.5.

3. ***Classification :*** Pour chaque concept considéré, un ou plusieurs classificateurs sont entraînés sur l'ensemble d'entraînement et les étiquettes associées, pour apprendre les paramètres de la fonction de prédiction et générer le modèle d'apprentissage. Plus de détails concernant les méthodes d'apprentissage sont donnés dans la section 2.6.

4. ***Fusion des descripteurs :*** la classification est effectuée séparément pour chaque classificateur et chaque descripteur. Les scores de prédiction de ces classificateurs individuels sont normalisés et ensuite fusionnés pour améliorer la fiabilité du système d'indexation. Plus de détails concernant les différentes stratégies de fusion sont donnés dans la section 2.7.

5. ***Fusion hiérarchique :*** Un autre type de fusion est appliquée suite à la fusion des descripteurs, la fusions hiérarchique. Les scores de classification obtenus pour différents types de descripteurs sont fusionnés par catégorie (par exemple les descripteurs de couleur ensemble, de texture ensemble et de mouvement ensemble).

6. ***Reclassement :*** ce post-traitement consiste à appliquer la méthode de reclassement temporel proposé dans [Safa 11b]. Cette méthode repose sur l'hypothèse que les vidéos ont statistiquement un contenu homogène, au moins au niveau local. Nous avons exploité cette hypothèse à une échelle globale et locale, en calculant un score de détection respectivement au niveau de la vidéo ou juste du voisinage de chaque plan puis en réévaluant le score de chaque plan en fonction de ce score.

Ce système peut être utilisé pour la détection de concept dans des séquences de vidéos ou dans les images fixes. Dans le cas des images fixes, seulement un sous-ensemble des descripteurs est utilisé (hors ceux liés au mouvement ou aux modalités audio) et la dernière étape de reclassement n'est pas appliquée.

2.4 Représentation des documents multimédias

La perception visuelle humaine est capable de reconnaître rapidement un très grand nombre d'objets, de visages et de scènes dans les vidéos même s'ils sont camouflés, de très petites tailles, et même si la personne est distraite ou occupée par une autre tâche [Li 02]. À l'opposé, pour une machine ces documents ne sont qu'un ensemble

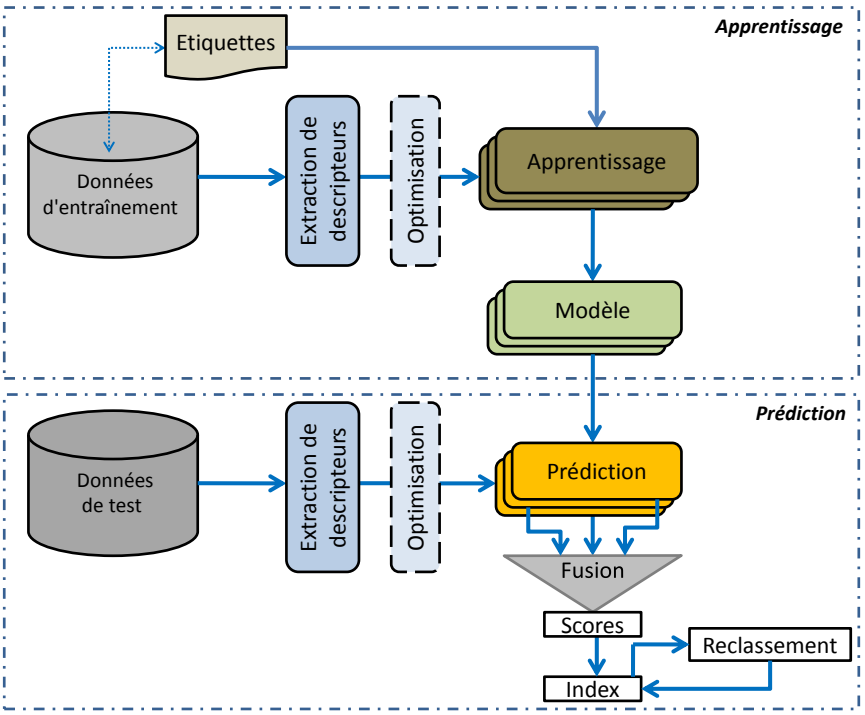


FIGURE 2.2 – Le processus général d'un système d'indexation classique.

d'information sous la forme de pixels et de d'échantillons sonores. Pour pouvoir procéder à une analyse automatique des documents multimédias, ces documents doivent être représentés sous la forme de descripteurs (le plus souvent des vecteurs dans $\mathbb{R}^d$) décrivant leur contenu visuel et audio des documents multimédias et assimilables par les machines. La représentation des documents multimédias est le processus de génération de descripteurs représentant d'une façon compacte l'information la plus significative contenue dans ces documents. Ces descripteurs sont également appelés caractéristiques, représentations, « features », attributs, ...

Le choix des descripteurs a une importance majeure au vu de son impact sur la qualité d'un système d'indexation automatique [Pari 10]. Un bon descripteur est le produit d'un certain équilibre entre quatre qualités : la robustesse, la capacité de discrimination, la performance en temps de calcul et la taille des vecteurs de description :

- robustesse : ils doivent être suffisamment robustes contre les variations d'éclairage, les artifices de compression des images, le flou et les déformations causées par un changement d'échelle ou par une rotation ;
- capacité de discrimination : Malgré la robustesse, les descripteurs doivent porter suffisamment d'information distinctives pour être capable de distinguer entre les différents concepts ;

- performance : l'extraction des descripteurs doit avoir des temps de calcul raisonnables. Cette caractéristique est primordiale, plus spécifiquement, pour les applications avec de contraintes temporelles comme le traitement de grands corpus ou le besoin d'un traitement en temps réel.
- taille : la taille des vecteurs de descriptions doit être minimale afin de respecter des contraintes d'espace de stockage et de temps de traitement.

Durant ces dernières années, le domaine de la génération de descripteurs a fait l'objet de nombreux travaux. Des descripteurs existent pour représenter différents types d'information à partir des vidéos comme la couleur, la texture, les formes, le son et le mouvement. Dans cette section, nous abordons les descripteurs les plus répandus pour la représentation du contenu des documents multimédias.

2.4.1 Descripteurs pour les images fixes

Deux grands types de descripteurs existent : les descripteurs locaux et les descripteurs globaux. Les descripteurs globaux représentent le contenu de l'image en entier alors que les descripteurs locaux représentent le contenu de certaines régions précises de l'image.

2.4.1.1 Descripteurs globaux

Les descripteurs globaux sont des descripteurs dont le temps de calcul est assez court. Ils sont relativement efficaces, ce qui les rend pratique pour les grandes collections de données. Pour les représentations globales d'images ou de vidéos il en existe de couleur, de texture ou de silhouette et de forme. Parmi les plus répandus et les plus utilisés dans le domaine de la classification visuelle on retrouve :

Couleur : Le descripteur couleur est probablement le descripteur le plus utilisé pour l'indexation automatique, il fournit une information forte et pertinente pour reconnaître des objets et discriminer les images. La plupart des images numériques sont représentées dans l'espace de couleur RGB même si d'autres espaces de couleur existent comme YUV, HSV, . . .Wan *et al.* ont étudié l'impact de différents espaces de couleurs RGB, YUV, HSV et CYLAB sur les performances d'indexation [Wan 98]. Le choix de l'espace de couleur joue un rôle très important dans la qualité du descripteur. En effet, ce descripteur dépend de deux aspects : l'espace de couleur choisi et sa représentation. En ce qui concerne la méthode de représentation de la couleur dans les images, celle la plus souvent utilisée est sous la forme d'histogramme de couleur, proposée pour la première fois par [Swai 91]. L'histogramme de couleur répartit les couleurs de l'image en plusieurs classes et compte ensuite le nombre de pixels appartenant à chacune des classes. Il décrit la distribution globale des couleurs dans une image sans aucune spécification des zones concernées. Ceci le rend résistant aux changements de point de vue et aux rotations d'objets dans les images, mais sensible aux changements d'échelle et d'éclairage [Geve 99]. Pour palier les problèmes

de déformations liées à la modification d'échelle, Huang *et al.* ont proposé d'ajouter une information spatiale à la distribution globale des couleurs de l'image, en ajoutant une corrélation spatiale entre paire de couleur [Huan 97]. Au lieu de décrire la distribution complète des couleurs de l'image, d'autres ont proposé de conserver la distribution des couleurs dominantes. Stricker *et al.* ont choisi de prendre en compte que les trois premiers moments statistiques de chacune des couleurs de l'espace couleur choisie [Stri 95]. Cette méthode permettrait de réduire les temps de calculs avec une précision comparable à celle obtenue avec les histogrammes de couleurs complets.

Texture : La texture de l'image s'est imposée comme une base visuelle importante pour l'indexation automatique de grandes collections d'images. Elle fournit des informations sur la surface des objets et les motifs visuels contenus dans l'image comme bois, métal, brique, velours. Trois méthodes principales sont utilisées pour l'analyse d'images, elles sont basées sur le filtre Gabor [Turn 86], sur des motifs locaux binaires [Ojal 96, Ojal 02] ou bien sur le fractal [Kapl 97]. Les méthodes les plus utilisées sont des méthodes issues du traitement de signal et basées sur les ondelettes et le filtre Gabor [Turn 86]. Ces méthodes capturent les fréquences et les directions principales dans l'image. Elles ont montré leur efficacité dans la tâche de classification et par rapport à d'autres méthodes utilisant des structures pyramidales (PWT) ou les modèles auto-régressifs multi-résolution (MR-SAR) [Mao 92, Manj 96].

Forme : Un des descripteurs globaux les plus répandus est celui proposé par Oliva *et al.*, le GIST [Oliv 01]. Ce descripteur représente la structure spatiale globale d'une image (scène) Au lieu de considérer une image comme une configuration de différents objets répartis dans l'espace spatial, Oliva *et al.* ont décidé de la considérer comme un seul objet avec une forme globale. Ce descripteur est efficace mais sensible aux transformations de l'image comme les coupures ou les rotations. Dalal et Triggs ont développé un histogramme de gradient orienté qui permet de caractériser la forme et l'apparence d'un objet avec une distribution de l'intensité locale du gradient (ou la direction des contours) sur une grille dense [Dala 05]. En pratique, ce descripteur (HOG) se calcule en divisant l'image en petites régions appelées cellules, et en calculant pour chaque cellule l'histogramme des directions du gradient ou des orientations des contours pour les pixels à l'intérieur de cette cellule (illustré dans la figure 2.3). La combinaison des histogrammes forme le descripteur HOG. Pour rendre ce descripteur invariant aux changements d'éclairage et d'ombrage, une normalisation en contraste peut être appliquée sur les histogrammes locaux. Ceci se fait en mesurant de l'intensité sur des zones plus larges que les cellules, appelées blocs, et en normalisant toutes les cellules du bloc par l'intermédiaire de cette valeur.

2.4.1.2 Descripteurs locaux

Les descripteurs globaux sont, en général, considérés comme des représentations sensibles aux fonds chargés et encombrés, et dépendant des changements de point de vue et des mouvements de caméra. Les descripteurs locaux quand à eux, ont

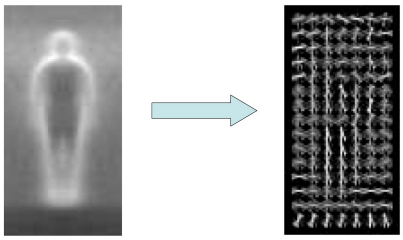


FIGURE 2.3 – Le descripteur HOG proposé par Dalal et Triggs pour la représentation de la forme d'un objet avec la distribution de l'intensité locale du gradient (ou la direction des contours) sur une grille dense [Dala 05]

été introduits pour fournir des représentations plus robustes et plus invariantes. Ils représentent des régions de l'image capables de caractériser l'image en question. Ces régions contiennent généralement un changement brusque des propriétés de l'image, par exemple : un changement de couleur, texture, de direction ou d'intensité. Biederman *et al.* ont montré l'importance de ces régions locales dans la reconnaissance d'objets même pour le système visuel humain [Bied 87]. Plus précisément, ils ont démontré expérimentalement que la suppression des coins ou des bords arrondis dans l'image gênerait la reconnaissance même humaine des objets contrairement à la suppression des bordures droites (voir figure 2.4).

Les descripteurs locaux sont calculés en deux étapes. La première étape consiste à extraire des régions ou des points d'intérêt caractérisant l'image (ou la vidéo), cette première étape s'effectue par l'intermédiaire d'un détecteur de régions locales. La deuxième étape consiste à décrire les régions sélectionnées de manière à représenter au mieux l'information contenue dans ces différentes régions.

Détecteur. Cette étape consiste à déterminer la localisation, la taille ainsi que le nombre de régions à extraire de l'image. Il existe trois méthodes principales pour la détection des ces régions : la détection de points d'intérêt, la détection de grille dense et l'échantillonnage aléatoire des régions.

Les détecteurs de points d'intérêt, comme leur nom l'indique, détectent les régions qui contiennent beaucoup d'information et qui peuvent être détectées précisément. Ces détecteurs ont l'avantage d'être en principe invariants aux transformations d'images géométriques (rotation, translation) et photométriques (variations d'intensité et de directions de l'éclairage). Une multitude de travaux a été effectuée dans ce domaine. Parmi les plus populaires, nous pouvons citer le détecteur de coin ou de points avec une forte courbure spatiale Harris [Harr 88] avec les extensions Harris-Laplace et Harris-Affine. Ainsi que le détecteur de « tâche » Hessian [Lind 98], MSER [Miko 04]. Lowe a introduit un détecteur de points d'intérêt, basé sur la

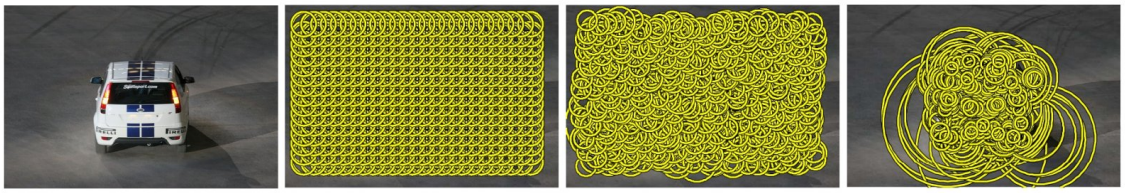
FIGURE 2.4 – L'importance des zones spécifiques de l'image comme les coins et les bords pour la reconnaissance des objets même pour le système visuel humain : la première colonne contient les objets complets, la deuxième colonne contient les mêmes objets qu'avec leurs coins arrondis (sans leurs bordures droites) alors que la troisième colonne contient les mêmes objets mais qu'avec leurs bordures droites et donc sans les bords arrondis [Bied 87]

différence de gaussienne, qui est invariant aux changements d'échelle, de translation, de rotation et au bruit dans l'image [Lowe 04].

La détection de grille dense est un échantillonnage dense qui sélectionne des régions selon une grille appliquée à l'image à différentes échelles pour assurer une invariance aux changements d'échelle. Cet échantillonnage dense permet de couvrir l'ensemble des objets contenus dans l'image ainsi qu'obtenir un nombre fixe de descripteurs pour chaque zone de l'image (voir figure 2.5). Nowak *et al.* ont montré que l'échantillonnage dense donne les meilleurs résultats dans le contexte de la reconnaissance d'objets [Nowa 06]. Tuytelaars *et al.* ont proposé un nouveau détecteur baptisé « points d'intérêt dense » qui combine les avantage de détecteurs de points d'intérêt à ceux de l'échantillonnage dense ainsi il permet d'améliorer les performances de la reconnaissance automatique [Tuyt 10].

Dans l'échantillonnage aléatoire, des régions de l'image sont sélectionnées aléatoirement [Nowa 06, Mare 05]. Ce détecteur ignore le contenu de l'image, de plus il ne possède pas d'avantages spécifiques comparé à l'échantillonnage dense.

Il existe un nouveau type de détecteur pour exploiter le mouvement sur une longue durée dans une vidéo, ce sont les détecteurs de trajectoires. Ces détecteurs suivent les régions d'une image tout le long d'images successives d'une vidéo. La



original image **dense sampling** **dense interest points** **interest points**

FIGURE 2.5 – Cette figure illustre la détection de régions d'intérêt selon trois différentes méthodes : l'échantillonnage dense, points d'intérêt dense et point d'intérêt [Tuyt 10]

plupart de ces détecteurs estime le flot optique de la vidéo pour retranscrire la trajectoire [Luca 81, Farn 03]. Ces détecteurs sont intéressants pour la reconnaissance d'action et d'événements dans les vidéos.

Descripteur. Une fois que les régions d'intérêt sont détectées, des descripteurs sont calculés afin de représenter l'information contenue dans ces régions. Comme les détecteurs locaux, les descripteurs locaux ont été largement étudiés. Là encore les descripteurs locaux doivent être invariants aux déformations d'images ou du moins robustes à celles-ci. Mikolajczyk *et al.* ont comparé plusieurs descripteurs locaux. Ils ont montré que ceux basés sur le SIFT (Scale Invariant Feature Transform) donnent les meilleurs résultats dans le contexte de la reconnaissance d'objets [Miko 05]. Ci-dessous nous abordons quelques descripteurs locaux parmi les plus populaires et les plus performants :

- SIFT (Scale Invariant Feature Transform) est le descripteur local le plus utilisé dans l'indexation d'images et de vidéos. Ce descripteur est invariant aux changements d'échelle, aux translations et rotations. Il est aussi partiellement invariant aux changements d'éclairage [Lowe 99, Lowe 04]. Ce descripteur calcule un histogramme d'orientation de gradient des différentes régions d'une image, comme illustré dans la figure 2.6. Pour construire le descripteur SIFT, Lowe propose de sélectionner une région de 4×4 pixels autour de chaque point d'intérêt détecté, de calculer ensuite un histogramme d'orientations de gradient par pixel en additionnant le nombre de gradient dans chacune des 8 orientations. Une fonction de pondération gaussienne est appliquée pour donner plus de poids aux gradients les plus proches du point d'intérêt considéré (représenté par le cercle bleu sur l'image). Enfin, les histogrammes des 8 orientations de gradient des régions de 4×4 pixels sont concaténés pour former le vecteur du descripteur SIFT de 128 dimensions ($4 \times 4 \times 8$) par région. Pour les approches utilisant différentes échelles, des régions plus ou moins grandes que 4×4 pixels peuvent être sélectionnées.

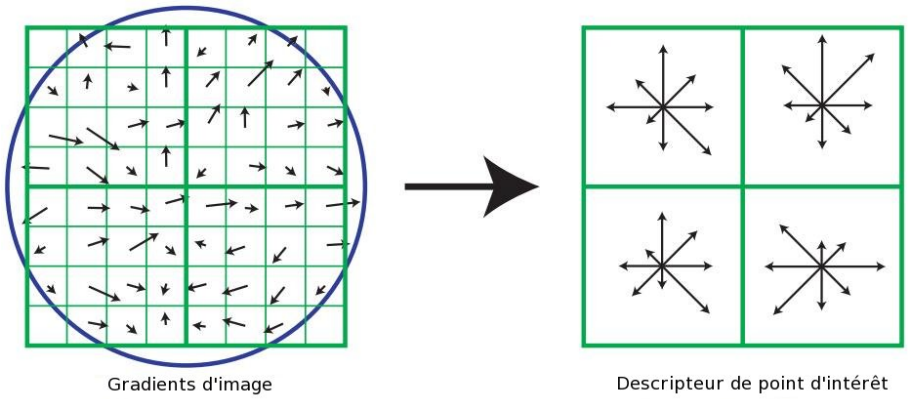


FIGURE 2.6 – Le calcul du descripteur SIFT se fait en sélectionnant une région de 4×4 pixels autour de chaque point d'intérêt détecté, en calculant ensuite un histogramme d'orientations de gradient par pixel par une addition du nombre de gradient dans chacune des 8 orientations [Lowe 04].

Depuis la mise en place du SIFT, un grand nombre d'extensions ont été proposées pour palier certains problèmes du descripteur SIFT tout en profitant des ses avantages. Le descripteur SIFT original a été conçu pour décrire les pixels d'une image en niveau de gris, il n'inclut donc aucune information sur les couleurs des régions d'intérêt. Nous trouvons une multitude d'extensions qui prennent en considération la couleur (Opponent-SIFT, W-SIFT, rgSIFT ou encore Transformed color SIFT [Geme 08], HSV-SIFT [Bosc 08], HueSIFT [Van 06]). Il existe aussi le RIFT, une extension pour rendre le SIFT invariant aux rotations [Laze 05] ou encore le MoSIFT pour représenter le mouvement.

- SURF (Speeded Up Robust Features) : est un descripteur invariant aux changements d'échelle et à la rotation, il est surtout connu pour sa rapidité (temps de calcul relativement court) et son efficacité [Bay 08]. Bay *et al.* ont proposé de décrire le point d'intérêt détecté et son entourage de 4×4 pixels par une distribution de Haar-wavelet responses.

2.4.1.3 Agrégation des descripteurs

Le nombre de descripteurs locaux extraits d'une vidéo est très variable. Le nombre de ces derniers peut exploser dans les grandes collections de données, ce qui entraîne des temps de calcul très longs et les rend inexploitables. Pour résoudre ce problème une étape d'agrégation est appliquée aux descripteurs calculés afin de capturer leur distributions statistiques. Cette agrégation génère une description plus globale à partir des descripteurs locaux, où différents niveaux de globalité sont possibles : au niveau de l'image, du plan ou de la vidéo entière. Dans ce qui suit, nous abordons les trois méthodes les plus populaires dans le domaine de la classification d'image

pour l'étape d'agrégation. Donc principalement la méthode de sacs-de-mots visuels, le noyau de Fisher et les VLAD.

Sacs-de-mots. Cette technique est inspirée des méthodes d'analyse de documents textuels. Le document texte est alors représenté par l'ensemble de mots qu'il contient sans prendre en considération leur ordre et leur structure [Harr 54]. La fréquence d'apparition de chaque mot dans le document est utilisée comme un descripteur du document texte pour la phase d'apprentissage. Cette méthode a été adaptée aux problématiques de l'analyse d'images par [Sivi 03, Csur 04] avec des sacs-de-mots-visuels (ou BoVW). Une image est alors représentée par un ensemble de sous parties de l'image, par exemple un visage est décrit par un ensemble ou un sac de sous parties (deux yeux, un nez, une bouche, ...) voir figure 2.7. La fréquence d'apparition de chaque sous partie (ou de chaque mot visuel) dans l'image est utilisée comme un descripteur de l'image pour la phase d'apprentissage.

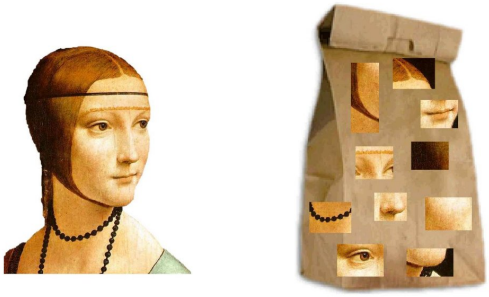


FIGURE 2.7 – Les sacs-de-mots visuels se basent sur le principe qu'une image peut être représentée par un ensemble de sous parties de l'image. Par exemple un visage est décrit par un ensemble ou un sac de sous parties [Fei 07] : deux yeux, un nez, une bouche, ...

Pour calculer la représentation en BoVW, il est tout d'abord nécessaire de trouver l'équivalent vidéo des mots des documents textes, les mots visuels, afin de créer le dictionnaire visuel. la représentation en BoVW se fait en trois grandes étapes (illustrées dans la figure 2.8) :

- L'extraction de descripteurs locaux : cette première étape consiste à extraire les descripteurs locaux de toutes les images de la collection de données.
- La génération du dictionnaire visuel : cette deuxième étape consiste à appliquer ensuite une méthode de regroupement (ou clustering) sur l'ensemble des descripteurs de la collection. La méthode de regroupement la plus populaire est la méthode des K-moyennes (ou K-means) : elle trouve des regroupements (ou clusters) de manière à minimiser la distance entre chaque descripteur local et son voisin le plus proche. Les centroïdes des regroupements constitueront

les mots visuels et l'ensemble des centroïdes formera le dictionnaire visuel de la collection.

- La représentation en sacs-de-mots-visuels (BoVW) : chaque descripteur local sera attribué au mot visuel (centroïdes) le plus proche en termes de distance. Dans le cas d'un BoVW flou, chaque descripteur local est attribué au centroïde le plus proche ainsi qu'aux centroïdes voisins relativement à la distance séparant le descripteur des différents centroïdes. Enfin, la représentation globale de l'image en BoVW est calculée sous la forme d'un histogramme additionnant la fréquence des mots visuels dans l'image.

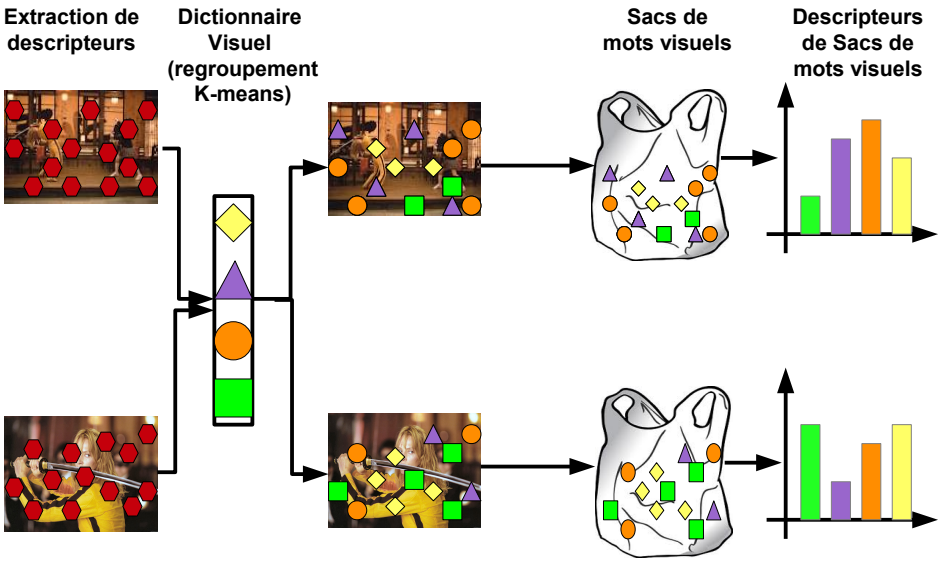


FIGURE 2.8 – Les différentes étapes pour la construction de la représentation des images en sacs-de-mots visuels.

Cette méthode est largement la méthode d'agrégation la plus utilisée et la plus étudiée de l'état de l'art. Cette popularité est due à la simplicité de sa mise en place et de son efficacité en termes de temps de calcul. Le descripteur obtenu par la représentation en BoVW, comme précisé précédemment, ne prend pas en considération les positions relatives des descripteurs locaux ce qui le rend invariant aux changements de point de vue et aux déformations globales. De plus, ce descripteur est plus robuste aux occlusions partielles. Ainsi la disparition de petits éléments de l'image ne l'impacte pas. Bien que la force de ce descripteur provienne de la non prise en compte des position relatives des descripteurs locaux (mots visuels), cette caractéristique constitue en même temps son principal inconvénient.

L'approche BoVW considère les images et les vidéos comme un ensemble d'éléments désordonnés, or dans la cadre de la classification d'images cette information spatiale est très importante. Pour palier cette absence d'information spatiale, Lazebnik *et al.* (Spatial Pyramid Matching) ont proposé de diviser l'image en grille rectangulaire de différentes résolutions et de calculer les histogrammes des BoVW pour chaque cellule de la grille [Laze 06]. Cette méthode a permis d'améliorer sensiblement les résultats de classification d'images. Des améliorations de cette représentation en sacs-de-mots-visuels continuent à être proposées [Laze 06, Liu 08, Quac 07, Geme 08, Yang 10, Van 10, Jego 10b, Lin 11, Jian 14].

Noyau de Fisher. La méthode de noyau de Fisher est une alternative à l'agrégation en sacs-de-mots. Alors que l'agrégation en sacs-de-mots se contente de compter les occurrences des mots visuels d'autre approches ont été proposées pour y introduire des statistiques plus avancées. Perronnin *et al.* ont proposé d'appliquer les noyaux de Fisher aux mots visuels dans le contexte de la classification d'images. Ils ont conçu les mots visuels avec un modèle de mélange de gaussienne (GMM) [Perr 07, Perr 10b]. Les deux méthodes d'agrégation, sacs-de-mots visuels et noyaux de Fisher, fournissent des résultats très comparables mais la représentation en noyau de Fisher génère des descripteurs de bien plus grandes dimensions pour une même taille de dictionnaire avec un temps de calcul inférieur à celui nécessaire pour le calcul des sacs-de-mots. Ceci rend la représentation en noyau de Fisher plus adapté aux très grandes collections de données (de 100 à 1 million de documents). Certaines méthodes proposent des approximations rapides des noyaux de Fisher [Jego 10a, Jego 12b, Delh 13].

2.4.2 Descripteurs vidéos

2.4.2.1 Descripteurs de mouvement

Les descripteurs locaux cités précédemment sont des descripteurs statiques qui ont fait leurs preuves dans le domaine de la classification d'images. Ces descripteurs ne prennent pas en compte la dimension temporelle ce qui limite leur capacité à représenter les vidéos.

Il existe une grande diversité de méthodes de représentations du mouvement dans les vidéos. Certaines méthodes utilisent la silhouette des personnes dans les images. Bobick *et al.* étaient parmi les premiers à utiliser cette méthode en extrayant les silhouettes d'une première image avec une soustraction de l'arrière plan et en agrégeant les différences entre la séquence d'images successives d'une action. Ceci donne une image binaire de l'énergie du mouvement (MEI) qui indique l'endroit où le mouvement se produit dans la vidéo et une image de l'histoire du mouvement (MHI) dont l'intensité des pixels est calculée à partir du mouvement de la silhouette (voir figure 2.9) [Bobi 01].

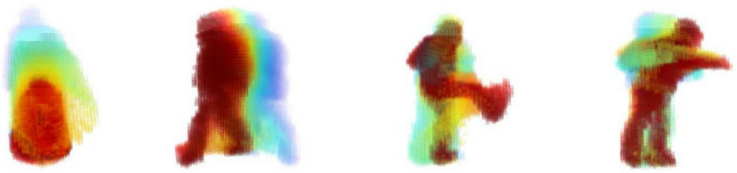


FIGURE 2.9 – La représentation du mouvement par une image de l'histoire du mouvement (MHI) où l'intensité des pixels est calculée à partir du mouvement de la silhouette agrégée sur une séquence d'images successives d'une action [Wein 06].

Dans le cas où plusieurs caméras sont utilisées, Weinland *et al.* ont proposé de combiner les silhouettes obtenues à partir de chacune des caméra en un seul voxel à trois dimensions (3D) [Wein 06]. Ils ont utilisé un volume de l'histoire du mouvement qui est une extension du MHI vers la 3D. Cette représentation est très informative mais nécessite une calibration très précise des caméra utilisées. Ces méthodes dépendent de la soustraction de l'arrière plan et ne tolèrent pas non plus les mouvements de caméras et les fonds dynamiques.

D'autres méthodes représentent le mouvement comme des volumes spatio-temporels à trois dimensions, formés par un empilage des images sur une séquence donnée [Batr 08, Blan 05, Gore 07, Yilm 08]. Pour représenter ces volumes spatio-temporels, Blank *et al.* ont d'abord empilé les silhouettes sur une séquence d'images précise pour former ces volumes (voir figure 2.10). Ils ont ensuite extrait des descripteurs spatio-temporels comme la saillance locale spatio-temporelle, la forme, la structure et l'orientation. Ces méthodes dépendent d'une soustraction du plan du fond et nécessitent une localisation et un alignement précis sur l'ensemble des images de la séquence donnée.

FIGURE 2.10 – La représentation du mouvement par un empilage de silhouettes sur une séquence d'images précise pour former des volumes spatio-temporels [Blan 05].

Dans le cas où la soustraction de l'arrière plan ne peut pas être appliquée, certaines approches ont choisi d'utiliser d'autres types informations comme la forme ou le mouvement. Par exemple Efros *et al.* ont calculé le flux optique (HOF) pour les images centrés sur la personne comme les vidéos de sports [Efro 03, Dala 06]. Cette méthode nécessite une segmentation et une stabilisation de chaque personne dans la séquence d'images de la vidéo.
Afin de réduire la sensibilité au bruit et aux changements de point de vue, certaines approches divisent l'image en grille spatiale ou temporelle où chaque cellule de la grille décrit localement une partie de l'image. Kellokumpu *et al.* ont calculé des motifs binaires locaux le long de la vidéo selon la dimension temporelle qu'ils ont stockés ensuite dans des histogrammes pour chacune des régions spatiales n'appartenant pas à l'arrière plan [Kell 08]. Alors que Thurau et Hlavac ont utilisé un histogramme des gradients orientés (HOG) en se focalisant sur les contours du premier plan [Thur 08]. Danafar *et al.* ont adapté les travaux d'Efros *et al.* pour générer une représentation du flux optique basée sur une grille horizontale divisant approximativement en tête, corps et jambes [Dana 07].

Une autre catégorie de méthodes existe pour la description du mouvement sous la forme d'une représentation locale basée sur des points d'intérêt spatio-temporels. Laptev *et al.* ont proposé d'étendre le détecteur de coin Harris vers la 3D (voir figure 2.11) pour retrouver les régions avec variations temporelles et spatiales significatives. Ils utilisent ensuite les HOF et HOG pour décrire ces régions [Lapt 05]. Dollar *et al.* ont apporté une amélioration à cette méthode en proposant un nouveau détecteur de points d'intérêt spatio-temporels plus adapté au contenu des vidéos [Doll 05]. Ces méthodes ont l'avantage d'être relativement invariantes au changement de point de vue et aux occlusions partielles et sans la nécessité d'une localisation ou de soustraction de plan de fond.

Enfin, une récente approche présentée par [Wang 11] a donné des résultats très intéressants dans la représentation du mouvement et la reconnaissance d'action. Cette méthode propose une représentation dense des trajectoires, inspirée par la technique d'échantillonnage dense habituellement utilisée pour l'extraction de descripteurs locaux statiques. Cette méthode est une extension des travaux effectués sur le calcul des contours des mouvements [Dala 06] sur les trajectoires denses. Elle combine donc la trajectoire, l'apparence et les informations concernant le mouvement. Le calcul de contours des mouvements sur les trajectoires denses rend cette représentation insensible aux mouvement de la camera.

2.4.2.2 Descripteurs audio

Les premiers travaux réalisés dans le domaine de l'indexation automatique des vidéos se sont surtout focalisés sur la représentation de leur contenu visuel. Les chercheurs ont pris conscience de l'importance de l'information audio, et ils incluent presque systématiquement les descripteurs audio dans leur système d'indexation des vidéos. En effet, pour la reconnaissance de certaines scènes comme « anniversaire » ou « ma-

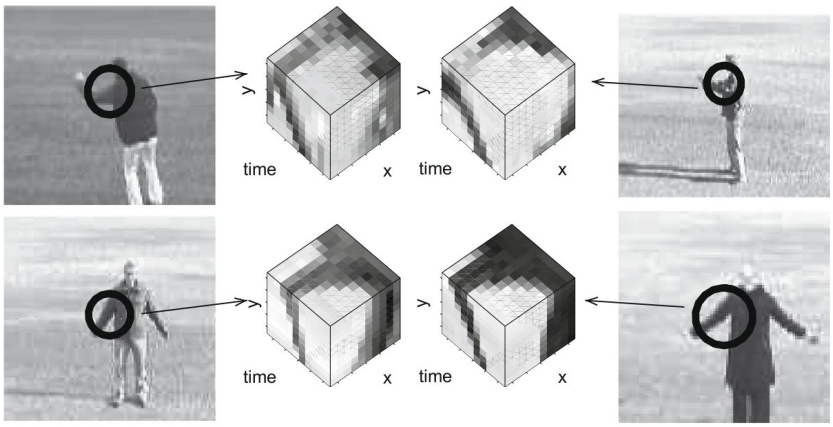


FIGURE 2.11 – La représentation du mouvement basée sur des points d'intérêt spatio-temporels. Ceci se fait pat l'intermédiaire d'une extension du détecteur de coin Harris vers la 3D pour retrouver des régions d'intérêt spatio-temporelles [Lapt 05].

riage », le contenu audio est bien plus discriminant et caractéristique que le contenu visuel.

Les descripteurs audio capturent des propriétés spécifiques au signal audio, comme la fréquence principale ou le volume du signal. Pour extraire les descripteurs audio, un pré-traitement et une segmentation du signal audio sont nécessaires car ce dernier est un signal unidimensionnel non stationnaire qui varie très rapidement dans le temps. Le signal audio est alors sous-échantillonné de façon à réduire la quantité de données à traiter tout en préservant son contenu spectral. Il est ensuite découpé en fenêtres d'analyse recouvrantes qui seront considérées lors du calcul des descripteurs. La durée de ces fenêtres est en général courte, variant entre 10 et 50 ms. Hormis la segmentation du signal audio en fenêtres, d'autres méthodes de segmentations sont également envisageables : par la détection du silence, de la musique, du bruit ou par l'identification du locuteur ou de l'émotion.

Mitrovic *et al.* présentent une étude relativement exhaustive des différents descripteurs audio utilisés dans différents domaines de recherche : reconnaissance automatique de la parole, la recherche d'information dans la musique, la segmentation audio, et la recherche de son environnemental [Mitr 10]. Ils ont repartis ces descripteurs en sept différentes catégories : temporels, fréquentiels, vectoriels, cepstraux, relationnels, modulation de fréquence, espace de phase. La boite à outil Spro[1] implémente les descripteurs audio classiques et elle met à disposition des utilisateurs une librairie en *C* pour pouvoir implémenter de nouveaux algorithmes.

[1] gforge.inria.fr/projects/spro/

Actuellement, le descripteur audio le plus utilisé est le MFCC (Mel Frequency Cepstrum Coefficients), un descripteur cepstral qui représente la densité spectrale de puissance du son. MFCC se base sur une transformée de Fourier rapide et d'une transformée de cosinus sur une échelle de Mel. Il se calcule de la manière suivante :

- Préaccentuation du signal pour relever les hautes fréquences.
- Découpage du signal en fenêtres d'analyse recouvrantes.
- Application de la transformée de Fourier rapide à la fenêtre pour obtenir le spectre.
- Application d'un banc de filtres triangulaires répartis sur l'échelle de Mel pour simuler l'oreille humaine.
- Application de la transformée de Cosinus discrète pour convertir le spectre logarithmique de Mel et représenter l'intensité du signal.
- Conservation des premiers coefficients pour former les coefficients cepstraux (MFCC).

À l'issue de cette phase d'extraction, les descripteurs sont agrégés de la même manière que les descripteurs visuels, par exemple sous la forme de sacs-de-mots audio ou par des GMM ou d'autres méthodes statistiques comme moyenne et variance ou minimum et maximum.

2.4.3 Descripteurs sémantiques

Les descripteurs précédents sont considérés comme des descripteurs de bas niveau, ils représentent le signal ou les caractéristiques basiques de l'image. Le descripteur sémantique, quant à lui, est considéré comme un descripteur de haut niveau. Il offre une représentation plus significative, pour la logique humaine, du contenu du document multimédia en modélisant la relation entre différents concepts apparaissant ou non dans un document multimédia. Smith *et al.* ont été les premiers à poser les bases d'une représentation sémantique, ils ont construit alors un vecteur de modèle qui combine les résultats d'un ensemble de modèles de concepts indépendants [Smit 03]. Chaque élément de ce vecteur de modèle correspond donc au score de confiance avec lequel chaque concept est détecté dans le document multimédia.

De nombreux travaux ont également été effectués sur la représentation sémantique [Jian 09, Li 10, Torr 10, Ayac 07a]. Ayache *et al.* ont conçu un modèle de numcepts mis en relation les uns avec les autres dans un réseau afin d'exploiter différentes formes de contexte et de permettre l'inférence ou la dérivation de nouveaux concepts [Ayac 07a]. Torresani *et al.* ont combiné les scores de détection obtenus par un grand nombre de classificateurs d'objets faiblement supervisés [Torr 10]. Li *et al.* ont ajouté à la représentation sémantique une notion de localisation des objets dans l'image

qui a permis d'améliorer considérablement les résultats du système de classification d'images [Li 10].

Cette représentation de l'image basée sur les objets a apporté des informations sémantiques complémentaires aux descripteurs de bas niveau, elle a ainsi prouvé son efficacité dans la classification des images. De plus, elle a ouvert une nouvelle direction de recherche dans le domaine de la classification des images et des vidéos. Récemment, des travaux ont adapté le descripteur sémantique pour la classification des événements dans les vidéos [Sada 12, Merl 12, Mazl 13, Habi 14a]. Merler *et al.* ont construit un descripteur sémantique à partir d'un très grand nombre de détecteurs d'objets (les 280 détecteurs de concepts à leur disposition) pour la détection d'événements dans les vidéos. En revanche, Mazloom *et al.* ont choisi de ne sélectionner que les détecteurs d'objets les plus informatifs pour un événement donné pour la construction de leur représentation sémantique [Mazl 13].

2.5 Optimisation des descripteurs

La conception d'un système d'indexation efficace, rapide et robuste est au cœur de toutes les recherches dans le domaine de l'indexation automatique. Alors que la majorité des travaux se concentrent sur l'élaboration de nouveaux descripteurs ou sur la mise en place d'un meilleur modèle d'apprentissage, peu s'intéressent à l'optimisation des descripteurs existant. Cependant, cette optimisation forme une étape importante dans le processus global de classification. Elle permet de former des vecteurs plus compacts et plus précis pour la représentation du contenu. De même elle diminue sensiblement le taux d'erreur des systèmes de classification et entraîne un gain dans les performances d'indexation. L'optimisation des descripteurs peut se faire de différentes façons : par une normalisation ou par une réduction de dimensions des descripteurs. Cette section explique les méthodes souvent utilisées pour améliorer les performances des descripteurs.

2.5.1 Normalisation des descripteurs

La normalisation des descripteurs a pour principal objectif de normaliser les composantes des vecteurs de description de façon à avoir une distribution uniforme. Parmi les méthodes de normalisation, nous citons celles qui s'appliquent au niveau des vecteurs séparément (par exemple la normalisation $l1$ ou $l2$), au niveau des composantes des vecteurs (par exemple la normalisation min-max ou variance unitaire) ou au niveau de chaque valeur indépendamment les unes des autres (par exemple la Power Law ou la normalisation de puissance). Dans ce qui suit, nous détaillons ces techniques. Soit V l'ensemble des N vecteurs de description à normaliser, chaque vecteur v_i est de dimension d (possède d composantes) $i = (1, 2, \ldots, d)$.

Normalisation $L1$ ou $L2$. Un grand nombre de descripteurs dépendent de la proportion de l'image qu'ils représentent, donc deux images contenant un même ob-

jet mais en différentes tailles auront deux descripteurs différents. Ces deux normalisations permettent de gommer la dépendance des descripteurs à la proportion de l'image.

Ces deux normalisations adaptent uniformément les composantes de chaque vecteur de façon à ce que la longueur du vecteur soit égale à 1 selon la métrique L_1 ou L_2.
Les vecteurs de description normalisés (V') sont donc produits en appliquant la formule suivante :

$$v'_{ij} = \frac{v_{ij}}{\|v_i\|}, i = 1, 2, \ldots, n, \text{ et } j = 1, 2, \ldots, d$$

où v_{ij} est la j^{ieme} composante du vecteur v_i, et $\|.\|$ est l'opérateur représentant la norme, qui est égale à $\sqrt{\sum_j v_{ij}^2}$ dans le cas de L_2normalisation ou égale à $|\sum_j v_{ij}|$ dans le cas de L_1normalisation.

Cette normalisation est adaptée pour les données qui respectent une distribution gaussienne et elle est la plus souvent utilisée pour normaliser les vecteurs de description sous la forme d'histogrammes et de sacs-de-mots (BoW). Néanmoins, elle est n'est pas restrictive et reste tout à fait applicable à d'autres types de descripteurs comme Perronnin *et. al* ont montré en l'appliquant sur les descripteurs en noyau de Fisher [Perr 10b].

Normalisation min-max. Cette normalisation vise à équilibrer l'influence des différentes composantes. Elle adapte les valeurs des vecteurs de description de façon à ce que toutes les valeurs soient comprises entre une borne inférieure et une borne supérieure (i, s).

Par conséquent, les valeurs du descripteur s'obtiennent par la formule suivante :

$$v'_{ij} = i + \frac{(s-i) \times (v_{ij} - min_k(v_{ij})}{max_k(v_{ij}) - min_k(v_{ij})}$$

Elle est souvent appliquée avec $i = 0$ et $l = 1$. Cette normalisation est actuellement utilisée dans libsvm [Chan 01].

Normalisation moyenne nulle et variance unitaire. Le but de cette normalisation est similaire à celui de la normalisation min-max, en revanche elle est moins sensible aux grandes variabilités entre les valeurs. Les valeurs des descripteurs sont normalisées en soustrayant la moyenne μ_j de chaque élément du descripteur et en divisant le résultat par la variance σ_j de l'élément du descripteur :

$$v'_{ij} = \frac{v_{ij} - \mu_j}{\sigma_j}$$

Où μ_j et σ_j sont la moyenne empirique et la variance de la j^{ieme} composante respectivement. Cette normalisation peut-être appliquée avec la partie variance unitaire uniquement.

Normalisation de puissance . (Power-law) Cette normalisation est surtout adaptée aux descripteurs sous la forme d'histogrammes. Elle concerne spécifiquement le cas où le nombre de regroupements (clusters) est relativement grand et donc où beaucoup de regroupements sont vides et ne contiendront aucun élément. Ceci entraîne des vecteurs à trous. Le but de cette méthode est de normaliser la distribution des valeurs des descripteurs en appliquant cette transformation à toutes les composantes des descripteurs individuellement [Safa 11a] :

$$x \leftarrow x^{\alpha}, x \leftarrow -(-x)^{\alpha}, \text{si } x < 0$$

Jegou *et al.* ont appliqué cette transformation sur des descripteurs de noyaux de Fisher avec $\alpha = 0.5$ et ils ont constaté empiriquement l'amélioration de la qualité de cette représentation [Jego 12b]. Ils ont montré qu'elle permet de stabiliser la variance et donc de corriger la dépendance entre la variance et la moyenne. Cette transformation est donc applicable aux descripteurs sous la forme de noyaux de Fisher comme aux sacs-de-mots [Perr 10b, Jego 12b, Winn 05a].

2.5.2 Réduction de dimensions des descripteurs

Le but de la réduction de dimensions est de garder les composantes les plus importantes seulement, c'est à dire celles qui portent le maximum d'information utile pour la classification. La réduction de dimensions permet alors d'obtenir un espace de descripteurs plus expressif pour cette classification. Ces méthodes statistiques ont été largement utilisées par la communauté du traitement d'image, de l'apprentissage automatique et en traitement du signal pour supprimer le bruit des données [La C 98, Hans 00, Kole 02, Vino 03]. Au sein de la communauté du multimédia, des travaux ont montré que les données visuelles et audio se prêtent bien à ce genre de traitement statistique pour mieux explorer les données [Fish 00, Smar 03]. Plusieurs techniques ont été proposées pour la réduction de dimensions des descripteurs pour l'indexation multimédia, comme la populaire méthode d'Analyse de Composantes Principales (ACP) [Joll 05] ou d'Analyse de Composantes Indépendantes (ACI) [Como 94].

Analyse de Composantes Principales (ACP). L'ACP propose une représentation dans un espace de dimension réduite, permettant ainsi de mettre en évidence d'éventuelles structures au sein des données. Pour cela, elle recherche les sous-espaces dans lesquels la projection des données déforme le moins possible les données initiales. Cette méthode proposée par Pearson [Pear 01] se fait par l'intermédiaire d'une projection linéaire des données originales dans un sous-espace de dimension inférieure à celle de l'espace de départ de façon à retrouver le maximum de variance possible entre les données. L'ACP calcule de nouvelles variables nommées composantes principales $(C_1, C_2, \ldots, C_k)$ qui sont obtenues par une combinaison linéaire des variables originales de la façon suivante :
Soit $(X_1, X_2, \ldots, X_p)$ les variables initiales dans l'espace p,

$$C_k = \alpha_{1k}X_1 + \alpha_{2k}X_2 + \cdots + \alpha_{pk}X_p$$

L'ACP permet d'extraire les informations les plus importantes des données, de compresser la taille des données en gardant seulement les informations importantes et de simplifier les descripteurs. De plus, elle s'est montrée utile pour enlever le bruit des données [Joll 05].

Analyse de Composante Indépendante (ACI). L'idée de l'ACI est similaire à celle de l'ACP. Les deux méthodes projettent les données dans différents espaces de composantes, à l'exception près que les composantes sont ici choisies de façon à ce qu'elles soient indépendantes. Le but des deux méthodes est différent : L'ACI transforme les données vers un espace de composantes indépendantes alors que l'ACP trouve les composantes non corrélés. L'ACI trouve les composantes statistiquement indépendantes et elle est préférée pour trouver les composantes les plus représentatives alors que l'ACP est idéale pour compresser les données dans un espace de dimension inférieure en supprimant les composantes les moins signifiantes.

Le fait que ces méthodes soient faciles à implémenter et que l'optimisation des descripteurs qu'elles permettent soit bien réelle les a rendu largement répandues dans le domaine de la classification. Cependant, elles ont une limitation commune à toutes les méthodes d'optimisation qui est leur linéarité. En effet ces méthodes sont toutes des méthodes linéaires alors qu'en pratique la distribution des données n'est pas linéaire. Pour palier ce problème de linéarité, une autre catégorie de technique de réduction de dimensions a été proposé, le « Local Embeddings »[Rowe 00]. Le point faible de ces derniers est leur sensibilité au bruit dans les données [Geng 05].

2.6 Méthodes de classification

Le processus de classification consiste à associer un élément à une classe donnée selon des caractéristiques précises. Les méthodes de classification sont basées sur des algorithmes d'apprentissage qui apprennent les corrélations entre les classes et les représentations de chacune des entités. En pratique, une fois que les descripteurs images (ou vidéo) sont extraits, des classificateurs sont entraînés sur l'ensemble de données appelé « ensemble d'entraînement » ou « ensemble d'apprentissage » pour apprendre les paramètres de la fonction de décision qui sépare les classes. Deux grandes catégories de méthodes de classification existent, les méthodes supervisées et les méthodes non supervisées.

Dans l'apprentissage supervisé, les étiquettes de l'ensemble d'entraînement sont connues, cet ensemble comprend des couples formés par l'élément (ou le descripteur) et l'étiquette correspondante (ou la classe). L'algorithme d'apprentissage analyse l'ensemble d'entraînement pour construire un modèle (ou une fonction de décision) qui permet de minimiser empiriquement l'erreur de classification dans cet ensemble. Dans l'apprentissage non supervisé, les étiquettes de l'ensemble d'entraînement ne sont pas connues. Le but est donc de trouver des structures cachées dans des données non étiquetées et la classification consiste à regrouper les éléments en classes non

nommées. Ces méthodes sont dites « méthodes de regroupement » (ou clustering), la méthode des K-moyennes (K-means) est probablement la plus connue dans cette catégorie.

Le choix de la méthode d'apprentissage automatique joue un rôle très important dans le système global d'indexation automatique et affecte directement sa qualité. Dans cette section, nous aborderons les méthodes les plus populaires en apprentissage automatique, en particulier la méthode des K-Plus Proches Voisins (ou KNN) et celle des Machines à Vecteurs de Support (ou SVM). Nous avons adopté ces deux méthodes d'apprentissage supervisé lors de nos expérimentations tout le long de cette thèse.

2.6.1 Les K plus proches voisins

L'algorithme des K plus proches voisins est un des algorithmes de classification les plus simples et les plus intuitifs. Il ne nécessite pas la construction d'un modèle d'apprentissage et son principe peut se résumer à « Dis moi qui sont tes amis, et je te dirais qui tu es ». Il consiste à prédire la classe d'un nouvel exemple en lui affectant simplement la classe majoritaire à partir des K exemples les plus proches [Cove 67]. À partir d'un ensemble d'apprentissage, d'une fonction de distance pour comparer deux exemples et d'un nombre de voisins K à prendre en considération, chaque nouvel exemple à étiqueter est comparé aux différents exemples de l'ensemble d'entraînement. Ensuite la classe de l'exemple est décidée par une combinaison linéaire des classes des K plus proches exemples pondérés par une fonction de distance au nouvel exemple. Plusieurs améliorations ont été apportées à cette méthode de classification, parmi elles nous citons la pondération des classes pour palier les problème de déséquilibre entre les classes.

Les performances de la méthode sont très dépendantes du choix de la fonction de distance intervenant dans le calcul des voisinages et du nombre de voisins K considéré. Malgré les nombreux travaux proposant un apprentissage de métriques dans le but de trouver la métrique de distance la plus appropriée aux données traitées [Wein 09, Chop 05], la distance Euclidienne reste la métrique de distance la plus utilisée dans l'implémentation des kNN. En ce qui concerne le paramètre K, il doit être choisi en fonction des données. Les grandes valeurs de K réduisent l'effet du bruit sur la classification mais entraînent un gommage des détails et rendent les frontières entre classes moins distinctes. À l'inverse les petites valeurs de K contiennent plus de variabilité et de sensibilité aux bruits mais rendent les frontières entre classes plus distinctes. Un bon choix de K peut se faire par différentes techniques heuristiques, par exemple par une validation croisée de façon à minimiser l'erreur de classification.

Bien que cette méthode de classification soit simple, efficace et contienne peu de paramètres, elle possède plusieurs limitations. Dans sa version originale cette méthode garde tous les exemples en mémoire, de ce fait elle nécessite beaucoup de mémoire. De plus, elle est très coûteuse en temps de classification vu qu'elle ne

construit pas de modèle d'apprentissage et que tous les calculs doivent être effectués lors de la classification. Pour ces raisons là, des méthodes d'optimisation de la gestion de l'espace mémoire et d'accès rapide sont inévitables pour être en mesure de l'utiliser sur les grands corpus.

2.6.2 Les machines à vecteurs de support

Les machines à vecteurs de support (ou SVM) sont les méthodes les plus répandues en apprentissage automatique pour la classification des données. Elles ont été introduites par [Cort 95] dans le cadre de la classification de documents textuels. Grâce à leur précision et leur fiabilité, elles ont vite été adoptées et sont devenues prédominantes dans la classification de documents multimédia également. Dans sa forme la plus simple pour la classification bi-classes, la méthode cherche à trouver l'hyperplan qui sépare « au mieux » les exemples d'apprentissage en deux classes. L'objectif est donc de trouver un séparateur linéaire qui maximise la marge entre l'hyperplan séparateur et les points dans la base d'apprentissage. Plusieurs extensions ont été proposées pour résoudre les problèmes des exemples qui ne sont pas linéairement séparables. La première relâche les contraintes de la marge dure pour accepter des solutions avec des exemples proches de la surface de séparation ou mal classés. La deuxième extension est basée sur l'utilisation de noyaux remplaçant le produit scalaire dans les calculs pour les exemples non linéairement séparables dans l'espace des caractéristiques choisi.

L'idée originale propose une séparation linéaire entre deux classes or les problèmes de classification souvent rencontrés comprennent bien plus que deux classes à reconnaître. De nombreuses stratégies ont été proposées pour étendre la classification binaire à différentes tâches d'apprentissage comme la classification multi-classes. Parmi les stratégies les plus simples, il existe celle qui entraine un classificateur binaire un-contre-tous pour chacune des classes sur l'ensemble des donnée d'entraînement. D'autres ont proposé des SVM multiclasses [West 99, Cram 02, Lee 04]. Joachims *et al.* ont proposé une alternative aux SVM multiclasses, en considérant le problème comme un problème de classement des exemples (ranking SVM) : parmi un ensemble de fonctions de classement, le modèle trouve la fonction qui maximise le gain. Grangier *et al.* ont amélioré le classement SVM (ranking SVM) en pondérant les classificateurs [Gran 08]. Akata *et al.* ont mené une étude comparative et ont montré que la stratégie simple un-contre-tous (ou OVR) surpasse les autres méthodes et les extensions multiclasse des SVM en termes de performance et de temps de calcul dans le cadre de la classification d'images [Akat 13]. En effet, la stratégie OVR est simple et rapide, elle décompose le problème d'apprentissage en problèmes indépendants par classe ce qui peut être parallélisable et donc plus rapide que les SVM multiclasse.

Trouver un apprentissage rapide et efficace pour les tâches de classification d'images reste un grand défi. Pour les petites collections de données de quelques milliers d'images, les classificateurs non-linéaires sont les plus utilisés grâce à leur

efficacité. Cependant, ces classificateurs non-linéaires ne sont pas adaptés pour les grandes collections de données de plusieurs millions d'images et plusieurs milliers de classes à cause de leur temps de classification très long. Le temps d'apprentissage des classificateurs précis peut atteindre jusqu'à plusieurs semaines voire des années selon Lin *et al.* [Lin 11]. Pour résoudre ce problème, beaucoup de travaux ont étudié les stratégies possibles pour améliorer la précision des méthodes d'apprentissage tout en réduisant les coûts très élevés des SVM non-linéaires. Certains travaux ont tenté d'optimiser les méthodes non-linéaires pour les rendre plus rapides en utilisant des algorithmes dits « optimization minimale séquentielle » (ou SMO) [Plat 99]. Ces algorithmes sont aujourd'hui implémentés dans la plupart des boîtes-à-outils : LibSVM [Chan 11], SVM light [Joac 99] et Shogun [Fran 08]. Toutefois le gain de temps de classification reste insuffisant pour permettre aux méthodes non-linéaires le passage à l'échelle pour les très grandes collections de données. D'autres travaux se sont focalisés sur la construction de classificateurs linéaires optimisés par l'utilisation des méthodes stochastiques de descente de gradient (ou SVM-SGD) proposée par LeCun *et al.* [LeCu 98, Bott 07, Shal 08]. Dans les travaux récents pour la classification d'images à grande échelle, l'optimisation par des méthode stochastiques de descente de gradient a eu beaucoup de succès et les méthodes l'utilisant se sont multipliées [Lin 11, West 10, Perr 10a, Rohr 11, Sanc 11].

2.6.3 Problème des classes déséquilibrées

Une autre problématique liée à l'apprentissage supervisé dans les grandes collections de données est le déséquilibre entre la classe positive et la classe négative qui affecte négativement la performance des classificateurs. Une solution simple est d'attribuer des poids plus élevés aux exemples des classes minoritaires [Domi 99, Elka 01]. Depuis ces travaux, un grand nombre d'algorithmes d'apprentissage (comme AdaBoost) ont été adaptés pour gérer les problèmes de déséquilibre entre classes par la pondération des exemples. Également, de nombreux algorithmes d'attribution de poids ont vu le jour [Ting 00]. La résolution du problème de déséquilibre des classes par cette méthode a montré de bons résultats dans le cas de faibles déséquilibres. Une alternative à la pondération des exemples consiste à modifier la taille des données d'entraînement. Elle se fait en sélectionnant aléatoirement un sous-ensemble à partir des données d'apprentissage majoritaires ou en dupliquant les exemples des données d'apprentissage minoritaire [Drum 03]. En plus des méthodes basiques d'échantillonnage, beaucoup de méthodes ont été proposées pour échantillonner d'une façon plus complexe et plus intelligente [Liu 09, Kuba 97, Chaw 11], ou encore Tahir *et al.* qui ont choisi d'inverser le sens du déséquilibre en formant des sous-ensembles à partir des données d'apprentissage contenant plus d'exemples de la classe minoritaire que d'exemples de la classe majoritaire [Tahi 09]. Ces méthodes ont été capables de palier les problèmes de grand déséquilibre [Zhou 06]. D'autres types de méthodes existent comme celles basées sur des techniques d'apprentissage, Safadi *et al.* ont prouvé l'efficacité des approches utilisant de multiples classificateurs (ou multi learner) pour les problèmes de déséquilibre [Safa 10] et ils l'ont combiné à

des algorithmes d'apprentissage actif pour améliorer leurs résultats [Safa 12].

Le système d'indexation multimédia utilisé tout au long de cette thèse adopte la méthode de classification proposée par [Safa 10], et basée sur de multiples classificateurs. L'idée principale de cette méthode de classification est de remplacer un grand classificateur déséquilibré par un certain nombre de plus petits classificateurs équilibré. La spécificité de cette approche réside dans la méthode de sous-échantillonnage des données d'apprentissage. En effet, pour un concept c, m sous-ensembles de données sont formés à partir des données d'apprentissage de la classe majoritaire selon la formule suivante :

$$m = \frac{f_{maj} \times nb_{maj}}{f_{min} \times nb_{min}}$$

Où nb_{maj} et nb_{min} sont des variables dépendantes des données d'apprentissage et qui représentent respectivement le nombre d'exemples de la classe majoritaire et de la classe minoritaire. f_{maj} et f_{min} sont des paramètres de contrôle. f_{maj} correspond au taux de recouvrement des exemples majoritaire, typiquement un $f_{maj} = 1$ signifie que tous les exemples de la classe majoritaire sont sélectionnés en moyenne au moins une fois dans les sous-ensembles. f_{min} correspond au ratio souhaité entre le nombre d'exemples de la classe majoritaire et celui de la classe minoritaire, et se fixe en fonction du déséquilibre que le classificateur tolère. Par exemple, pour un classificateur SVM f_{min} se situe idéalement entre 2 et 4.

Les m sous-ensembles de données d'apprentissage contiennent tous les exemples de la classe minoritaire (nb_{min}) et $f_{min} \times nb_{min}$ d'exemples sélectionnés aléatoirement des données d'apprentissage de la classe majoritaire. Un classificateur est entraîné sur chacun des sous-ensembles de données d'apprentissage et génère le modèle d'apprentissage correspondant. Ensuite, les m modèles d'apprentissage obtenus sont utilisés sur les données de test pour prédire la probabilité d'apparition (ou encore le score de prédiction) du concept c. Enfin, les m scores de prédiction résultants sont fusionnés pour générer un seul score final pour chacun des exemples des données de test.

2.7 Fusion

La fusion d'information intervient dans les systèmes possédant plusieurs sources d'information (visuelles, textuelles, audio, ...). Cette fusion de différentes sources d'information constitue une étape primordiale pour les systèmes d'indexation. Elle peut être effectuée à différents niveaux : au niveau de la représentation (fusion précoce), au niveau de la décision (fusion tardive) ou à un niveau intermédiaire (fusion des noyaux). Nous avons illustré ces trois niveaux de fusion dans la figure 2.12. Altrey *et al.* ont étudié la plupart des méthodes de fusions multimodales proposées durant ces dernières années afin d'apporter une meilleure compréhension et

un meilleur choix en fonction de la problématique [Atre 10].

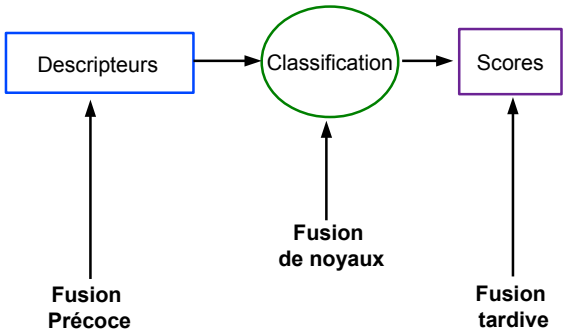


FIGURE 2.12 – Les trois différents niveaux de fusion de plusieurs sources d'information dans les systèmes d'indexation : au niveau de la représentation (fusion précoce), au niveau de la décision (fusion tardive) ou à un niveau intermédiaire (fusion des noyaux)

2.7.1 Fusion précoce

Dans la fusion précoce les différents descripteurs sont combinés avant la classification [Gong 08]. La simplicité d'implémentation de cette fusion (concaténation des descripteurs des différentes modalités) l'a rendu populaire dans le domaine. Cependant, il est souvent problématique de transformer différents descripteurs en une seule représentation. En effet chaque descripteur provient de différents espaces avec différentes distributions, ainsi une grande différence d'échelle des valeurs ou du nombre de dimensions des descripteurs concaténés peut entraîner un déséquilibre dans la prise en compte de certains descripteurs par rapport aux autres. Dans ces cas, il sera nécessaire de normaliser ou de pondérer les différents descripteurs après la concaténation. Également, la concaténation de plusieurs descripteurs peut générer des descripteurs à très grandes dimensions et entraîner par la suite un temps d'apprentissage beaucoup plus long. Dans ces cas il sera nécessaire de réduire la dimension du descripteur final en appliquant une méthode de réduction de dimensions, par exemple une analyse en composantes principales (ACP).

2.7.2 Fusion tardive

Dans la fusion tardive les scores de classification obtenus séparément par chacun des modèles de descripteur sont combinés [Snoe 05, Derb 12, Gian 10, Lin 09]. Contrai-

rement à la fusion précoce, la fusion tardive s'appuie sur la force de chacune des modalités séparément. L'avantage de cette fusion est sa flexibilité dans la mesure où il est possible d'utiliser la méthode de classification la plus appropriée à chacune des modalités (celle qui considère au mieux la spécificité de la modalité). De plus, la combinaison de différentes méthodes de classification permet de pallier les erreurs de prédiction de chacune séparément et fournit souvent des décisions plus précises. Néanmoins, ce gain de précision a un coût qui se traduit en une augmentation de temps de calcul vu que chaque modalité nécessite sa propre étape d'apprentissage, en plus de la perte de corrélation entre les modalités. Dans le but de profiter des avantages de chacune des méthodes de fusion, des approches hybrides proposent de combiner les deux niveaux de fusion, précoce et tardive, dans le but de profiter des avantages spécifiques à chacune d'elles [Lan 13].

2.7.3 Fusion de noyaux

La fusion de noyaux peut être considérée comme une fusion intermédiaire entre la fusion précoce et la fusion tardive. La fusion de noyaux combine les modalités au niveau du noyau. Au lieu de modéliser les données selon une seule fonction de noyau (comme dans la fusion précoce), cette fusion offre la possibilité de choisir le noyau le plus adapté à chacune des modalités et de combiner ensuite les noyaux uni-modaux pour générer un seul noyau final multi-modal [Ayac 07b, Muhl 12]. Elle permet d'exploiter le maximum d'information de chacune des modalités. L'inconvénient de cette méthode de fusion est le nombre de paramètres à fixer d'abord sur l'ensemble des fonctions de noyaux pour chacune des modalités, et ensuite sur la fonction de fusion pour le noyau final.

Finalement, le principal problème des méthodes de fusion de l'état de l'art est la perte potentielle de corrélation entre les modalités dans un espace contenant plusieurs descripteurs, plus on va vers la fusion tardive plus on perd les corrélations vu que la fusion se fait plus loin dans le signal. Dans le Chapitre 3, nous proposons une nouvelle méthode de fusion, intitulée « fusion doublement précoce », pour palier les problèmes de perte de corrélations inter modalités.

2.8 Apprentissage profond ou *Deep Learning*

Les méthodes de l'état de l'art pour l'indexation de contenu multimédia sont basées sur l'apprentissage en surface, c'est-à-dire sur l'application de méthodes d'apprentissage supervisé sur des descripteurs extraits selon des méthodes déterministes pensées par les chercheurs ou *ad hoc*. Ces méthodes ont montré leur efficacité d'indexation sur les données bien structurées et propres. Néanmoins leur capacité limitée de modélisation et de représentation pose problème dans l'indexation des données difficiles à conditions réelles. En effet, le succès des systèmes d'indexation dépend de la représentation de données alors que ces données peuvent être très variables à cause de

certains facteurs appelées « facteurs de variations » : des changements d'orientation, d'éclairage, de positions. Des connaissances physiques de chacun des facteurs de variations sont nécessaires afin de rendre les systèmes robustes en intégrant des formules mathématiques modélisant les variations. Or la plupart des facteurs de variations contenus dans les images naturelles n'ont pas de caractéristiques modélisables.

L'apprentissage profond (*Deep Learning*) peut apporter une solution à ces problèmes de représentation. Cet apprentissage permet d'extraire des structures complexes et de construire des représentations internes à partir de données variées. Son but est de découvrir automatiquement des abstractions directement à partir des descripteurs bas niveau jusqu'à des concepts de haut niveau. Il tente d'apprendre une hiérarchie des représentations de manière à ce que les représentations d'un niveau soient formées par une composition des représentations de niveau inférieur [Beng 09]. L'apprentissage profond est composé de multiples couches de transformations non-linéaires contrairement à l'apprentissage en surface qui ne comprend qu'une ou au maximum deux couches de transformations non-linéaires. La profondeur de l'architecture correspond au nombre de niveaux de la composition d'opérations non linéaires de la fonction apprise.

Les systèmes d'indexation basés sur l'apprentissage profond ont longtemps été mis de côté dans le domaine de l'indexation multimédias à cause de leur sensibilité aux annotations bruitées et incomplètes et surtout de leur très long temps de calcul. Grâce à la création de très grands corpus correctement annotées (par exemple ImageNet) et à la disponibilité de nouvelles architectures (parallélismes, GPU, ...), les systèmes basés sur l'apprentissage profond trouvent de plus en plus de succès dans l'indexation de contenu multimédias. Plus précisément, une classe de méthodes d'apprentissage profond, les réseaux de neurones convolutionnels (ou Convolutional Neural Networks), est en vogue actuellement pour l'indexation des images [LeCu 10, Kriz 12] et très récemment pour l'indexation des vidéos [Wang 11, Fara 13].

L'architecture d'un réseau de neurones convolutionnels, par exemple, en 8 couches se compose de trois couches de réseaux de neurones complètement inter connectés (*all-to-all*) précédées de 5 couches à convolution qui permettent d'avoir un traitement invariant par la translation et de réduire le nombre de coefficients à considérer (comme illustré dans la figure 2.13).

Ces systèmes possèdent une contrainte au niveau de l'image d'entrée qui doit être de taille relativement petite ce qui exige un redimensionnement des images et une sélection de différentes zones de l'image (par exemple les 4 coins et la partie centrale de l'image et leur inverse [Jia 13]). Jia a mis en place un système libre et optimisé implémentant les algorithmes des réseaux de neurones convolutionnels [Jia 13, Kriz 12], ce qui rend ces méthodes accessibles au public et utilisables sur les grands corpus d'images et de vidéos.

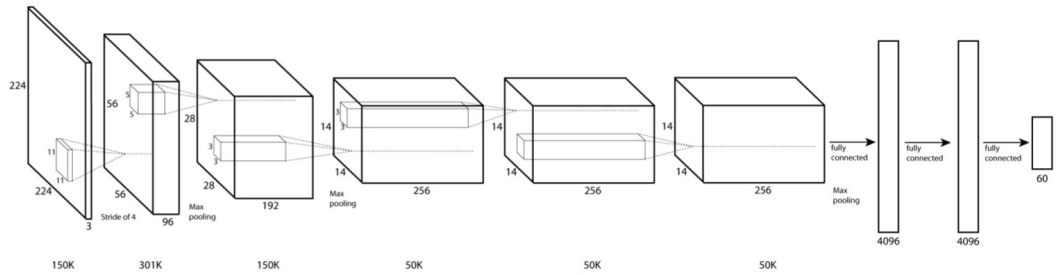


FIGURE 2.13 – Une illustration de l'architecture d'un réseau de neurones convolutionnels (CNN) en 8 couches [Kriz 12, Snoe 13].

Enfin, la mise en application de l'apprentissage profond étant relativement récente de nombreuses perspectives d'améliorations ainsi que de nouvelles méthodes d'indexation sont à prévoir sur le court terme.

2.9 Collections de données

Pour être en mesure d'évaluer et de tester l'indexation automatique multimédia, plusieurs collections de données ont été crées. Ces collections, le plus souvent publiques, forment une base commune permettant de comparer différents systèmes d'indexation selon les mêmes métriques d'évaluations. Au total plus de 100 collections sont disponibles entre images de divers objets, de logo, de fleurs, d'animaux ainsi que des vidéos de mouvements, d'actions humaines, de sports, de caméra de surveillance, de films ou de séries télévisées.

Parmi les collections les plus utilisées pour l'indexation des images fixes nous trouvons Voc, ImageNet et Caltech. En revanche, pour l'indexation des vidéos nous citons les collections suivantes : la collection de données Hollywood2 pour la reconnaissances des actions dans les films, MediaEval pour la détection de la violence dans les films, HMDB pour la détection et reconnaissance des actions dans les vidéos et enfin la grande collection de données fournie par la campagne d'évaluation annuelle TRECVid pour la reconnaissance d'un grand nombre de concepts statiques et dynamiques (événements) dans les vidéos du web. Chaque collection possède sa propre spécificité et complexité, le tableau 2.1 récapitule leurs principales caractéristiques.

2.9.1 KTH

La collection de données KTH comprends 2 391 vidéos enregistrées spécifiquement pour une validation expérimentale scientifique [Schu 04]. Les vidéos contiennent 6 catégories d'actions humaines : *boxing, hanclapping, handwaving, jogging, running, walking*. Les enregistrements sont effectués par 25 différentes personnes dans 4

Collection	Année	Taille	Concepts	Action	Type	Difficulté
KTH	2004	2 392 vidéos	6	Oui	Encadrées	-
Hollywood2	2009	1 684 vidéos	13	Oui	Films	++
PASCAl VOC	2010	11 540 images	30	Non	Web	+
HMDB	2011	6 766 vidéos	51	Oui	Web	++
MediaEval-VSD	2013	18 vidéos	2	Oui	Films	++
TRECVid-MED 2013	2013	98 000 vidéos	20	Oui	Web	++
TRECVid-SIN 2013	2013	800 000 vidéos	364	Oui	Web	++

TABLE 2.1 – Les principales caractéristiques des collections de données les plus populaires figurent dans ce tableau. Parmi ces caractéristiques, nous trouvons : la taille de la collection, le nombre de concepts statiques et dynamiques qui y sont annotées, si la collection comporte certaines action humaines, le type des données (films, tirées d'internet ou spécifiquement crées pour les évaluations) enfin le niveau de difficulté des données (arrière-plans encombrés, mouvements de caméra, ...).

différents contextes : à l'extérieur, à l'extérieur avec différentes échelles, à l'extérieur avec différents vêtements et à l'intérieur. Les vidéos filmées respectent un certain nombre de contraintes : un arrière-plan homogène et gris, une personne par séquence vidéo, pas de mouvement de caméra, pas de changement de point de vue. Pour cela nous appellerons ces vidéos des « vidéos encadrées ».

2.9.2 PASCAL Visua Object Classes (Voc)

PASCAL VOC, est parmi les plus grandes collections d'images annotées publiques [Ever 10, Ever 11]. Elle contient 11 540 images collectées du site de partage de photos en ligne Flickr avec des annotations de 20 objets (*person, bird, car, airplane, table, plant,* ...) et 10 actions (sauter, marcher, lire, courir, ...). Cette collection de données offre la possibilité d'évaluer les méthodes de reconnaissance d'objets dans une très large palette d'images et dans des scènes réelles. Pour cela, les images choisies contiennent une très grande variabilité en terme de taille d'objets, d'orientation, de pose, d'éclairage, de position et d'occlusion. Une annotation exhaustive et précise des images a été faite pour assurer une base d'entraînement propre et une évaluation précise des méthodes de reconnaissance d'objets.

En parallèle, une campagne d'évaluation annuelle (VOC) basée sur ces données a été mise en place entre 2007 et 2012. Son principal objectif est de fournir des images difficiles et des annotations de bonne qualité avec un système d'évaluation standard pour permettre à différents algorithmes de se comparer et de mesurer les capacités des méthodes de l'état de l'art. Quatre tâches ont été définies :

- la classification : prédire la présence ou non des objets dans les images,
- la détection : prédire l'emplacement des objets dans les images avec un cadre englobant (voir figure 2.14),

- la segmentation : prédire pour chaque pixel de l'image à quel objet il apparient,
- la classification d'actions : prédire pour chaque action si une personne l'effectue dans les images et si c'est le cas une boite doit être dessinée pour localiser la personne.

FIGURE 2.14 – Un exemple des objets à localiser dans les images selon la tâche de « détection » de VOC.

2.9.3 Hollywood2

Cette collection contient des séquences courtes extraites de 69 films de l'industrie cinématographique (par exemple *American Beauty, As Good As It Gets, Being John Malkovich, Big Fish, . . .*) [Mars 09]. Ces séquences sont des fragments correspondant à des plans de films. Tous les fragments ont été annotés selon huit différentes actions (voir figure 2.15) : *answer phone, drive car, eat, kiss, run, sit down, . . .* et dix scènes dont deux sont filmées à l'extérieur et huit à l'intérieur : *house, Bedroom, kitchen, car, restaurant, office, shop, . . .*. La collection contient au total 3 669 plans annotés pour un équivalent de 20,1 heures de vidéos.

Le but de cette collection est de fournir des données réalistes et complexes pour l'évaluation de systèmes de reconnaissance d'actions humaines. Grâce aux conditions expérimentales très variées et contraignantes, les vidéos de la collection Hollywood2 sont plus difficiles que d'autres collections de données représentant des actions humaines (comme KTH).

FIGURE 2.15 – Quelques exemples d'actions parmi les huit actions annotées dans Hollywood : « Kiss », « answer phone », « get out car », « hug ».

2.9.4 HMDB

Cette collection comporte 51 types d'actions avec au moins 101 vidéos par actions pour un total de 6 766 vidéos. Les vidéos ont été collectées à partir de plusieurs sources et elles ont annotées manuellement [Kueh 11]. HMDB permet d'évaluer la

qualité et la robustesse d'un système de reconnaissance d'action grâce aux multiples conditions d'enregistrement des vidéos, comme les mouvements de caméra et de point de vue, les occlusions ainsi que la qualité des vidéos.
Pour capturer au maximum la richesse et la complexité des actions humaines, les vidéos ont été collectées d'internet, Youtube, Google, des films et des archives comme Prelinger. La figure 2.16 montre quelques exemples d'actions parmi les 51 actions contenues dans cette collection. Ces 51 actions ont été regroupées en cinq grandes catégories : les actions faciales (*smile, talk, laugh, ...*), les actions avec la manipulation d'objets (*smoke, eat, drink, ...*), les mouvements du corps (*run, jump, walk, climb, dive, ...*), les mouvements de corps avec manipulation d'objets (*hit something, golf, ride bike, ride horse, ...*) et enfin les mouvements de corps pour des interactions humaines (*fencing, hug, kiss, ...*).

FIGURE 2.16 – Quelques exemples d'actions parmi les 51 actions annotées dans HMDB : « hand wave », « drink », « fight », « jump », « run »

2.9.5 MediaEval

MediaEval a créé un référentiel multimédia au sein de la communauté de recherche qui s'intéresse aux aspects sociaux et humains des données. Chaque année MediaEval lance des défis à la communauté multimédia en proposant des tâches variées et les données correspondantes pour permettre d'évaluer et comparer les méthodes existantes. Chaque année les mêmes tâches sont proposées avec un certain nombre de nouvelles tâches.

Parmi ces tâches nous trouvons celle de recherche de photos sociales diverses qui consiste à affiner une liste classée de photos de lieux extraites de Flickr en utilisant

des informations visuelles et textuelles fournies sur un ensemble de 95 000 photos et 200 lieux. Il y a également la tâche de recherche et d'hyperliens, cette tâche demande de trouver les segments de vidéos pertinents pour une requête donnée et de fournir une liste d'hyperliens utiles pour chacun des segments. Ou encore la tâche de reconnaissance d'émotion dans la musique consiste à détecter automatiquement les émotions dans la musique en utilisant le contenu et générer une représentation continue. Les données fournies contiennent 744 chansons.

Enfin, la tâche de détection de violence (VSD) consiste à détecter les portions des films contenant de la violence en utilisant les informations multimodales des films (audio, visuel, textuel). Une trentaine de films de l'industrie cinématographique ont été annotés, en plus d'un ensemble de vidéos courtes d'Internet (de Youtube ou d'archive d'Internet) nécessaires pour l'extension de la tâche sur la détection de violence dans les vidéos courtes [Dema 13]. Dans cette thèse, nous nous sommes intéressés à cette dernière tâche en proposant un modèle corrélant plusieurs types d'information pour la détection de la violence dans les vidéos (voir Chapitre 3).

2.9.6 TRECVid

La campagne d'évaluation TREC Video (TRECVid) a commencé à petite échelle en 2001, motivée par l'intérêt du NIST à étendre la recherche d'information au delà du texte et par l'absence de bases communes pour la comparaison des résultats de recherche dans les vidéos [Over 13]. Depuis TRECVid a continué à s'agrandir pour former une campagne d'évaluation annuelle incontournable dans le domaine de le recherche d'information dans les documents multimédia à grande échelle. Les données et les annotations sont fournies par les organisateurs mais elles ne sont pas toutes publiques. La campagne comprend plusieurs tâches d'indexation de documents multimédias. Dans cette thèse, nous nous sommes particulièrement interessés à la tâche de Semantic Indexing (SIN) pour la reconnaissance d'objets (*dog, airplane, boat,* ...) ou de scènes (*classroom, harbor,* ...) et la tâche de Multimedia Event Detection (MED) pour la détection des événements comme *wedding, birthday,* ....

Les tâches de TRECVid sont considérés comme des tâches très difficiles à traiter pour plusieurs raisons. Premièrement, le très grand nombre de concepts à détecter (346 différents concepts entres objets, scènes ou actions simples et 20 événements) oblige la mise en place d'un système de détection générique de la part des participants. La quantité de données considérablement élevée nécessite des systèmes performant en temps de calcul et d'un certain nombre de ressources. La collection contient plus de 98 000 vidéos et 4 600 heures pour la détection des événements et 800 000 plans courts sur 400 heures de vidéo pour la détection des concepts.

La grande variabilité des données en termes de provenance (vidéos professionnelles ou amatrices), de qualité (filmé avec une caméra de qualité ou avec un téléphone) et de contenu (avec des changements d'éclairage, de fond, de mouvement de caméra ou des différents environnement). De plus, la fréquence des concepts et des

événements dans les vidéos n'est pas uniforme, certains apparaissent bien plus souvent que d'autres. Enfin, les annotations fournies par les organisateurs sont faites au niveau des plan en entier sans aucune précision sur l'endroit et le moment exact de leur apparition dans les vidéo. Ainsi, il suffit que le concept (ou l'événement) apparaisse dans une image du plan pour que ce dernier soit annoté positif pour le concept en question. Ceci constitue une sérieuse difficulté pour les modèles d'apprentissage qui doivent reconnaître lesquelles des informations extraites du plan sont représentatives ou non du concept (ou événement). Nous nous sommes basés sur cette collection de données pour évaluer nos travaux tout au long de ce mémoire (voir Chapitres 4, 5 et 6).

2.10 Conclusion

Pour conclure ce chapitre, nous soulignons que toutes les étapes du système d'indexation par le contenu, détaillées précédemment, sont importantes et affectent directement leur performances. Nous détaillons les limitations identifiées et qui ont justifié les directions de recherche choisies pour ce travail de thèse.

Nous nous intéressons à la détection de concepts statiques et dynamiques dans les vidéos difficiles en conditions réelles (à l'opposé de celles qui sont encadrées ou jouées). La représentation du contenu multimédia est une étape clé du processus d'indexation. Les descripteurs doivent être capables de représenter le contenu multimodal des vidéos en conservant le maximum de corrélations intermodalités. Or le principal problème des méthodes de fusion de l'état de l'art est la perte potentielle de ces corrélations. Par conséquent nous proposons une nouvelle méthode de fusion multimodale pour générer une représentation multimodale conjointe du contenu vidéo (Chapitre 3).
Les concepts complexes ou événements sont généralement composés d'une association d'un ensemble de concepts basiques apparaissant en même temps, qu'il est important de localiser. Pour cela nous présentons, dans le Chapitre 4, une méthode faiblement supervisée de localisation de concepts, comme des objets dans les images, qui sera utile pour détecter ensuite des concepts plus complexes (comme des événements).
En ce qui concerne l'apprentissage supervisé, les données d'apprentissage doivent être complètes et précises mais les annotations sont souvent effectuées au niveau de la vidéo. Dans de nombreux cas, des annotations plus fines, au niveau du plan, sont nécessaires. Nous explorons la possibilité de classifier automatiquement les plans de vidéos à partir d'annotations globales niveau vidéos dans le Chapitre 5.
Enfin, les descripteurs doivent également être optimisés de façon à améliorer la capacité de ceux-ci à représenter le contenu multimédia tout en réduisant leur coût en termes de temps de calcul et de stockage. Pour cela, nous exposons une méthode d'optimisation de descripteurs dans le Chapitre 6.

3

Motifs audio-visuels joints

Dans ce chapitre, nous présentons une nouvelle méthode de fusion intitulée fusion « doublement précoce », pour palier les problèmes de perte de corrélations entre les modalités. Par l'intermédiaire de cette fusion, nous proposons une représentation audio-visuelle des données pour détecter les événements dans les vidéos. Ainsi, nous proposons un descripteur qui fournit des indices multimodaux audio et visuels ; tout d'abord en assemblant les descripteurs audio et visuels, ensuite en révélant statistiquement les motifs conjoints multimodaux.
Nous avons évalué la performance de ce nouveau descripteur issu de la fusion « doublement précoce » proposée dans le contexte de la détection de scène violentes dans les vidéos. Plus précisément, la validation expérimentale a été effectuée dans le cadre de la tâche de détection de scènes violentes de la campagne d'évaluation MediaEval 2013.

3.1 Motivations : Exploitation de la corrélation entre différentes modalités

Aujourd'hui, des millions de documents multimédias sont crées et partagées quotidiennement par des professionnels, des simples utilisateurs amateurs ou par l'industrie cinématographique. Rien n'est mieux que quelques chiffres pour mesurer l'ampleur du volume des données multimédias accessibles. Plus de 100 heures de vidéos sont téléversées chaque minute sur la plateforme de partage en ligne Youtube en 2013. Et plus de 550 millions d'images sont téléversées au quotidien en 2013 selon le rapport annuel de Mary Meeker. Selon l'Institut de Statistique de l'UNESCO (ISU) la production mondiale annuelle de films a atteint 6 573 longs métrages en 2011. Devant la quantité écrasante de documents multimédias, la classification manuelle est devenue impossible et un grand besoin d'automatisation se fait sentir pour faciliter l'accès aux contenus multimédias.

Les actions et les événements dans les vidéos se déroulent sous la forme de scènes visuelles souvent accompagnées par des informations audio spécifiques (par exemple pour un événement comme « anniversaire » les vidéos comportent des images de ballons, de gâteaux et d'un grand nombre de personnes souvent accompagnées par une musique festive et plus précisément de la chanson spécifique aux anniversaires). Ces informations visuelles et audio forment donc des motifs conjoints audio-visuels. Ainsi une méthode efficace pour la détection des événements doit exploiter les deux modalités : audio et visuelle. La technique la plus populaire d'analyse audio-visuelle est basée sur une fusion multimodale. Or, le principal problème des méthodes de fusion de l'état de l'art (voir 2.7) est la perte potentielle de corrélation entre les modalités dans un espace de descripteur mélangés. Plus on va vers la fusion tardive, plus on perd les corrélations vu que la fusion se fait plus loin dans le signal. Même si la fusion précoce reste l'approche capable de respecter au maximum les inter-corrélations, les approches de fusion précoce de l'état de l'art représentent séparément le contenu des vidéos en concaténant n histogrammes monodimensionnels correspondant aux n

modalités considérées.

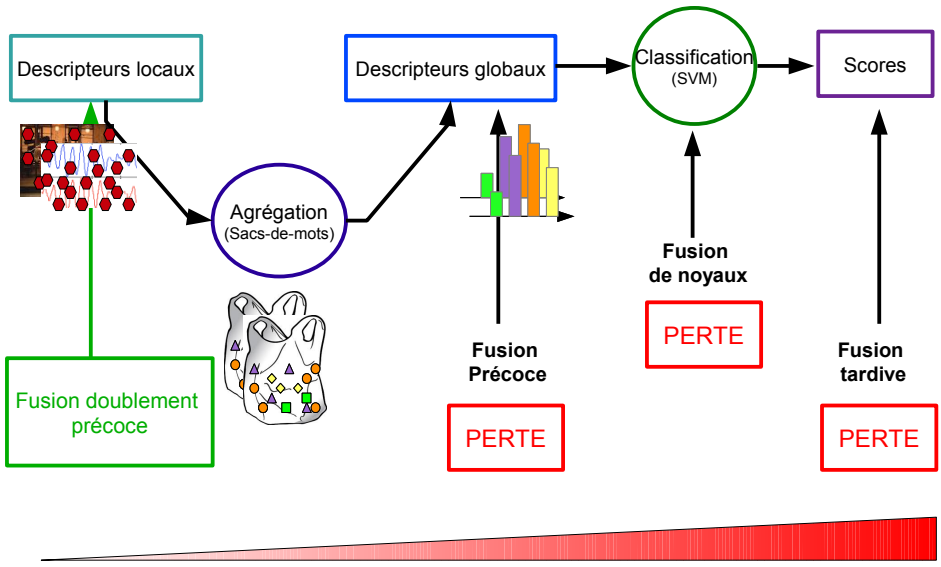


FIGURE 3.1 – La nouvelle méthode de fusion proposée dans ce chapitre (fusion doublement précoce) et les trois différents niveaux de fusion de plusieurs sources d'information dans les systèmes d'indexation : au niveau des descripteurs globaux obtenus par une agrégation par exemple sous la forme de sacs-de mots (fusion précoce), au niveau des noyaux de la méthode d'apprentissage par exemple un SVM (fusion de noyaux), au niveau des scores de prédiction (fusion tardive).

En effet, les histogrammes multidimensionnels fournissent, habituellement, une représentation plus fine du contenu que celle de plusieurs histogrammes monodimensionnels. Par exemple dans le cas des images, l'histogramme tridimensionnel de couleur RGB est plus discriminant que trois histogrammes monodimensionnels (R, G et B). Pour illustrer cette différence, nous considérons deux images dont la première contient du rouge et du bleu et dont la deuxième contient du noir et du violet. Bien que ces deux images soient visuellement très différentes, les trois histogrammes monodimensionnnels (R, G et B) en donnent exactement la même représentation. L'histogramme tridimensionnel RGB est par contre en mesure d'en fournir deux représentations effectivement très différentes. La figure 3.2 illustre cet exemple. Notons que les méthodes de fusion par noyaux et tardives prennent encore moins en compte la corrélation entre les éléments des différentes modalités puisque la fusion se fait encore plus loin du signal.

Nous proposons ici une méthode de fusion (qu'on appelera une fusion « doublement précoce ») qui permet de décrire le contenu de la vidéo en se basant sur la

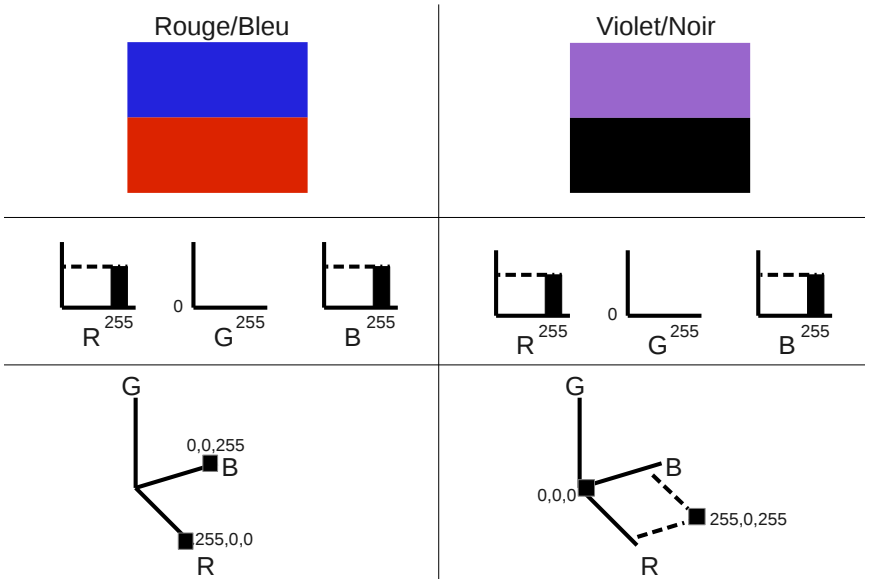


FIGURE 3.2 – Les représentations fournies par un histogramme tridimensionnel RGB versus celles fournies par trois histogrammes monodimensionnels (R, G et B) pour deux images distinctes.

relation conjointe entre les deux modalités : la modalité audio et la modalité visuelle (figure 3.1). Notre approche est basée sur une représentation sous forme des sacs-de-mots audio-visuels et est évaluée dans le cadre de la détection d'un événement spécifique : la détection de scènes violentes dans les vidéos.

3.2 Travaux connexes

Plusieurs techniques ont été mises en place et implémentées pour l'analyse du contenu audio et du contenu visuel des vidéos conjointement. Pour la détection des événements dans les vidéos, Ye *et al.* ont modélisé la relation entre la modalité audio et la modalité visuelle avec un graphe bipartite suivie d'un partitionnement de ce graphe de façon à révéler des motifs joints [Ye 12]. Dans le domaine de reconnaissance des événements dans les vidéos de surveillance, certains ont proposé des méthodes pour intégrer les informations audio et visuelles [Cris 07]. Ils calculent une matrice de co-occurrence audio-visuelle pour détecter et segmenter les événements sous la forme de données audio-visuelles. Dans le domaine du suivi d'objets, Beal *et al.* ont décidé d'exploiter la structure statistique des données sonores et visuelles ainsi que leurs dépendances mutuelles. Ils l'ont traduite dans un seul modèle graphique probabiliste [Beal 03]. Sargun *et al.* ont utilisé l'Analyse Canonique des Corrélations (ACC) pour fusionner l'information sonore et l'information visuelle et créer un espace multimodal dans le but d'améliorer la performance des systèmes audio-visuels de reconnaissance du locuteur. L'Analyse Canonique des Corrélations (ACC) permet de trouver deux espaces basiques dans lequel la matrice de corrélation croisée

entre les variables est diagonale et les corrélations diagonales sont maximisées. Par l'intermédiaire de l'ACC, ils ont fusionné le descripteur audio et celui de la texture des lèvres en concaténant les composantes corrélées des vecteurs de description audio et visuels [Sarg 06, Sarg 07]. Enfin, dans le domaine général de classification de concepts dans les vidéos, Jiang *et al.* ont étudié la causalité temporelle statistique entre les mots audio et visuels pour représenter le contenu de la vidéo comme étant des motifs audio-visuels [Jian 11].

Dans le domaine de la détection de scènes violentes dans les vidéos, la littérature ne propose pas de définition générale de la violence. Sachant que définir le terme « Violence » n'est pas une tâche facile à cause de son ambiguïté et de sa subjectivité, chaque scientifique a dû clarifier sa propre description de la violence. On peut trouver des définitions littéraires comme : « violence physique ou accident amenant à des blessures humaines ou de la douleur » [Dema 13]. Il existe aussi des définitions plus techniques où la violence est définie par des indicateurs visuels et audio spécifiques, par exemple les mouvements accélérés ou les rythmes de musique rapides [Gong 08]. Les travaux dans ce domaine sont souvent basés sur des descripteurs visuels ou spatio-temporels [Datt 02, Berm 11, Souz 10]. D'autres méthodes se focalisent sur l'unique utilisation des descripteurs audio [Gian 06]. Certains proposent même une nouvelle utilisation de la représentation en sacs-de-mots audio classiques, en décrivant chaque segment par un ou plusieurs mots audio obtenus par une simple agrégation sur les descripteurs audio classiques [Pene 13].
Dans l'ensemble, les systèmes de détection de violence sont limités et globalement moins évolués que ceux de l'indexation d'événements à cause du manque de définitions communes de la violence et de corpus de vidéos commun.

3.3 Descripteur audio-visuel proposé

Cette section décrit la représentation audio-visuelle jointe que nous proposons pour la détection des événements et plus précisément de scènes violentes. Le but étant d'exploiter la forte corrélation entre l'information audio et l'information visuelle afin de découvrir des motifs audio-visuels capables d'identifier les scènes violentes. La représentation des motifs audio-visuels est censée donner de meilleurs résultats qu'une simple fusion (précoce ou tardive) des deux modalités audio et visuelle qui ignore leurs corrélations. La méthode proposée est composée de trois étapes :

1. Dans un premier temps, les descripteurs locaux audio et visuels sont extraits à partir de la vidéo ;
2. Ensuite, les motifs bimodaux (ou encore les mots bimodaux) sont trouvés et le dictionnaire bimodal est construit ;
3. Enfin, la représentation sous la forme de sacs-de-mots bimodaux est construite par l'intermédiaire de ces mots.

Le processus général de la méthode est illustré dans la figure 5.2.

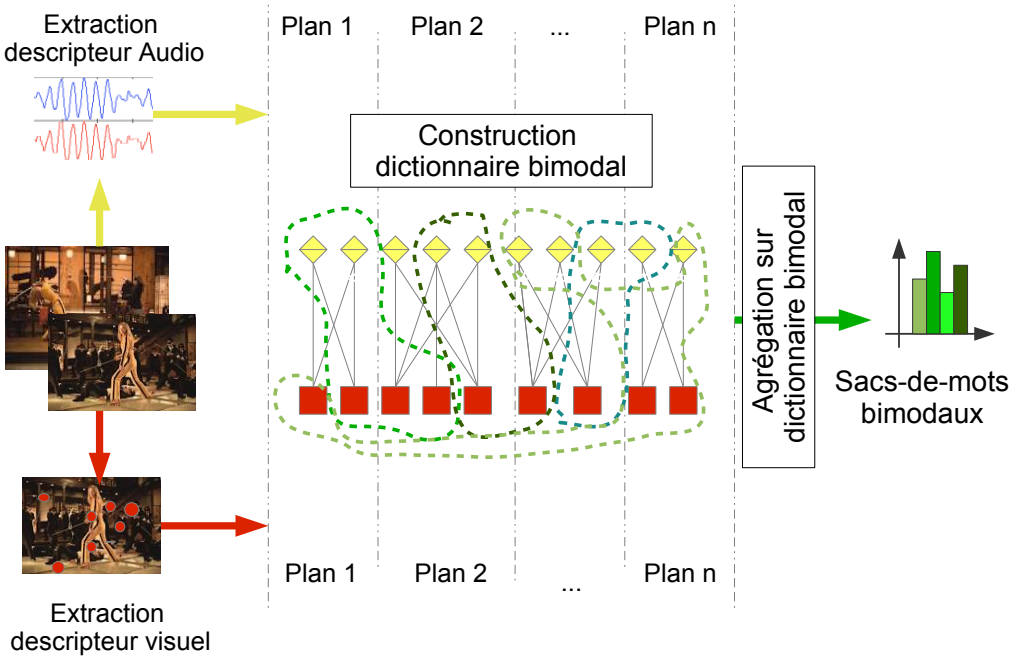


FIGURE 3.3 – Le processus général pour la génération de sacs-de-mots audio-visuels

3.3.1 Extraction des descripteurs locaux

On considère une collection de vidéos décomposée en n_s plans vidéo. Les descripteurs locaux audio et visuels sont extraits pour chacun des plans. Pour chaque plan vidéo s_i, ceci peut être écrit sous la forme de a_{ij} et v_{ik}. Tous les a_{ij} (resp. v_{ik}) sont des vecteurs de dimensions fixées à d_a (resp. d_v) et le nombre de descripteurs locaux n_{ai} et n_{vi} dépend généralement du contenu du plan vidéo s_i. Pour chaque plan vidéo s_i, ceci peut être noté comme suit :

- Descripteurs audio locaux a_{ij} : un certain nombre de descripteurs audio sont extraits (par exemple des MFCC).
- Descripteurs visuels locaux v_{ik} : un certain nombre de descripteurs visuels sont extraits (par exemple des descripteurs de points d'intérêt SIFT).

Où

- $1 \leq i \leq n_s$, $1 \leq j \leq n_{ai}$ et $1 \leq k \leq n_{vi}$,
- n_{ai} est le nombre de descripteurs audio locaux dans le plan vidéo s_i,
- n_{vi} est le nombre de descripteurs visuels locaux dans le plan vidéo s_i.

3.3.2 Capture de motifs bimodaux

Dans l'approche classique de sac-de-mots et par modalité, a_{ij} (resp. v_{ik}) sont agrégés sous la forme d'histogramme selon les groupes (clusters) calculés précédemment sur la totalité des descripteurs locaux du même type disponibles sur la collection de données. Les groupes (clusters) constituent alors un dictionnaire de taille fixe prédéfinie qui correspond également à la taille de représentation agrégée.

Pour la représentation audio-visuelle jointe, l'ensemble de vecteurs m_{ijk} est défini comme suit :

$$m_{ijk} = \{a_{ij} \otimes v_{ik}\} \text{ avec } 1 \leq i \leq n_s, 1 \leq j \leq n_{ai} \text{ et } 1 \leq k \leq n_{vi}$$

$a_{ij} \otimes v_{ik}$ est simplement le produit tensoriel de a_{ij} et v_{ik}. Tous les m_{ijk} doivent avoir une même dimension fixe $d_a + d_v$ et le nombre de descripteurs locaux $n_{avi} = n_{ai} \times n_{vi}$ dépend généralement du contenu du plan vidéo s_i.

Avant l'application du produit tensoriel sur les descripteurs audio et visuels, une normalisation et éventuellement une pondération peuvent être appliquées. La normalisation peut être faite par un coefficient multiplicatif global de façon à ce que la distance moyenne entre deux descripteurs locaux soit égale à 1 pour ramener les descripteurs à échelle équivalente. En ce qui concerne la pondération, celle-ci peut être effectuée selon la performance relative des descripteurs audio et visuels considérés séparément et évalués par une validation croisée sur l'ensemble d'apprentissage.

Le nombre de descripteurs audio peut être élevé, de même que le nombre de descripteurs visuels pour chacun des plans vidéo de la collection. De plus, le grand nombre de plans vidéos dans la collection de données peut entraîner la génération d'un très grand nombre de descripteurs audio-visuels joints. Même si la représentation n'aura pas à être stockée le processus d'agrégation doit lui être appliqué, d'où le besoin de trouver une solution pour réduire leur nombre. Dans le but de rendre l'approche généralisable et applicable dans le cas où plus de deux descripteurs locaux auront à être fusionnés de cette manière, nous proposons de limiter le nombre de combinaisons audio-visuelles considérées pour un plan donné à une certaine valeur seuil n_{max}. Dans ce cas, la représentation locale audio-visuelle jointe sera un ensemble de m_{il} avec n_{ml} étant le nombre de descripteurs locaux audio-visuels dans le plan vidéo s_i :

$$\{m_{il}\} \subseteq \{m_{ijk}\} \text{ avec } 1 \leq i \leq n_s \text{ et } 1 \leq l \leq n_{ml} \\ 1 \leq j \leq n_{ai} \text{ et } 1 \leq k \leq n_{vi}$$

Si le nombre total de combinaisons audio-visuelles générées ($n_{ai} \times n_{vi}$) est inférieur ou égal au seuil fixé par plan (n_{max}), tous les vecteurs audio-visuels ($a_{ij} \otimes v_{ik}$) seront considérés. Alors que si le nombre de combinaisons audio-visuelles générées ($n_{ai} \times n_{vi}$) est supérieur à ce seuil (n_{max}), seulement n_{max} vecteurs seront sélectionnés aléatoirement à partir de l'ensemble de vecteurs audio-visuels m_{ijk}.

Ceci peut être décrit comme suit :

$$n_{ml} = \begin{cases} n_{ai} \times n_{vi} & \text{si } n_{ai} \times n_{vi} \leq n_{max} \\ n_{max} & \text{sinon} \end{cases}$$

On note que le nombre de vecteurs maximal à prendre en considération n_{max} est fixé par validation croisée sur l'ensemble d'apprentissage.

Ensuite, une méthode standard de regroupement sera appliquée sur les n_{max} vecteurs joints pour capturer la corrélation entre l'information audio et visuelle et donc retrouver les motifs audio-visuels.

3.3.3 Représentation sous la forme de sacs-de-mots bimodaux

Enfin, l'agrégation de type sacs-de-mots peut être appliquée sur l'ensemble d'apprentissage exactement de la même manière sur les m_{il} descripteurs locaux que sur les a_{ij} et v_{ik}.

3.4 Evaluations

3.4.1 MediaEval2013

L'efficacité de notre représentation audio-visuelle jointe a été mesurée dans le cadre de la tâche de détection de scènes violentes de MediaEval2013. Cette tâche définit deux types de violence : violence objective et violence subjective. La violence objective est définie comme étant « violence physique ou accident résultant en blessures humaines ou douleur ». La violence subjective est définie comme étant « les scènes qu'on ne pourra pas laisser un enfant de 8 ans regarder dans un film à cause de la violence physique qu'elles contiennent » [Dema 13]. La figure 3.4 montre quelques images extraites des plans annotés comme contenant des scènes violentes objectives.

FIGURE 3.4 – Quelques images extraites des plans annotés comme étant violents

La collection de données comprend 25 films hollywoodiens décomposés en 43 923 plans et répartis en deux ensembles : l'ensemble d'apprentissage et l'ensemble de test. L'ensemble d'apprentissage contient 18 films annotés décomposés en 32 678 plans. L'ensemble de test contient 7 autres films hollywoodiens décomposés en 11 245 plans. Les données de l'ensemble d'apprentissage sont annotées par plan comme contenant ou non des scènes violentes (subjective ou objective) en plus de dix autres concepts : *sang, feu, cris, poursuite de voitures, arme à feu, gore, arme blanche, explosions, coups de feu, combats*. Ces dix concepts peuvent être utilisés par les participants pour détecter les scènes violentes. Le nombre de plans annotés comme étant violents ou non sur l'ensemble d'apprentissage et de test est récapitulé dans le tableau 3.1. La tâche de la détection de scènes violentes constitue un réel défi à cause de la difficulté de la représentation de la violence et de la rareté des échantillons positifs (scènes violentes) dans l'ensemble des vidéos (8.28% de la durée totale pour la violence objective et 13.91% de la durée totale pour la violence subjective).

	Objective			Subjective		
Nombre de plans	App	Test	Total	App	Test	Total
Violents	3 921	1 180	5 101	7 010	2 276	9 286
Non violents	28 757	10 065	38 822	25 668	8 969	34 637
Total	32 678	11 245	43 923	32 678	11 245	43 923

TABLE 3.1 – Le nombre de plans annotés comme étant violents ou non sur l'ensemble d'apprentissage et de test de la collection de données de MediaEval 2013.

Pour choisir et fixer les paramètres de notre modèle, nous procédons par validation croisée. Nous découpons la collection d'apprentissage en deux sous-collections de taille équivalente qui serviront de base pour la validation croisée.

3.4.2 Choix de paramètres

Pour générer le descripteur audio-visuel joint proposé, nous utilisons un descripteur audio classique « Mel Frequency Cepstral Coefficients » (MFCC) pour représenter le contenu audio de la vidéo. Les points d'intérêt spatio-temporels (STIP) sont quant à eux utilisés pour représenter l'information visuelle de la vidéo [Lapt 05] vu que le mouvement est très important pour la détection de violence et que ce descripteur met l'accent sur le mouvement. Avant la génération de la représentation audio-visuelle jointe, les deux descripteurs ont été optimisés séparément sur les 12 concepts annotés (fournis par les organisateurs de la tâche de détection de scènes violentes de MediaEval 2013) pour éviter le sur-apprentissage sur les deux concepts cibles (violence objective et violence subjective). Cette optimisation sera détaillée dans ce qui suit. En ce qui concerne la classification, elle a été effectuée sur deux méthodes d'apprentissage différentes, la première basée sur des SVM multiples (MSVM) [Safa 10] et la deuxième basée sur la recherche de K plus proches voisins.

L'outil d'Ivan Laptev[1] a été utilisé pour calculer les points d'intérêt spatio-temporels (STIP) [Lapt 05]. Un vecteur d'histogramme de flux optique (HOF) est ainsi produit pour chaque STIP détecté. La dimension de ce vecteur est de 90 éléments. Une agrégation est ensuite appliquée sur la durée du plan. Nous notons que le descripteur composé d'histogramme de gradients (HOG) a été essayé également mais il n'a pas fourni de meilleurs résultats que le descripteur HOF, de même que leur fusion (HOF et HOG).

L'outil de Guillaume Gravier Spro[2] a été utilisé pour le calcul du descripteur audio MFCC. Un vecteur dont la dimension est de 13 éléments est produit par le programme chaque 10 ms. Nous avons comparé la performance du descripteur MFCC seul (13 éléments) à celle du descripteur MFCC incluant les coefficients delta et accélération sur les données d'apprentissage. Habituellement, ces coefficients (delta et accélération) permettent d'améliorer la performance du descripteur MFCC, mais ceci n'a pas été vérifié lors de nos expérimentations. Le coefficient d'accélération ne serait probablement pas adapté à la représentation des bruits comme des explosions qui sont pertinents ici. Pour la suite de nos expérimentations, nous nous sommes contenté alors du descripteur MFCC seul (13 éléments).

La durée minimale de la fenêtre impacte directement la performance de ce descripteur audio. Nous avons optimisé cette durée minimale par validation croisée sur l'ensemble d'apprentissage. La figure 3.5 montre la performance des MFCC en fonction de différentes durées de fenêtre.

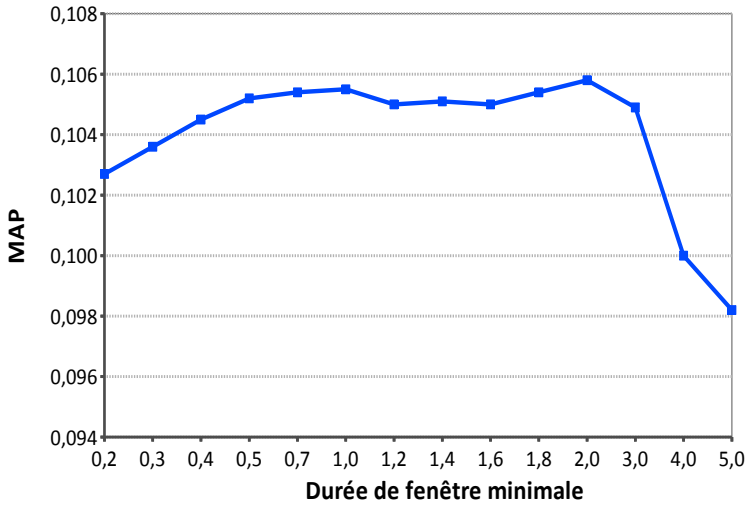


FIGURE 3.5 – La performance des MFCC en fonction de la durée minimale de la fenêtre.

[1] http://www.di.ens.fr/~laptev/download.html
[2] http://www.irisa.fr/metiss/guig/spro/

Une durée minimale de 1.8 secondes donne les meilleures performances. Une agrégation est appliquée sur la durée du plan étendue avant et/ou après pour atteindre 1.8 secondes si celle-ci est inférieure à cette valeur.

De plus, le nombre de groupes (clusters) calculé sur les descripteurs locaux extraits (taille du dictionnaire) et utilisé pour la génération de la représentation en sacs-de-mots influence sensiblement la performance de ces descripteurs. Nous avons également optimisé la taille du dictionnaire par une validation croisée sur l'ensemble d'apprentissage. L'influence du nombre de groupes sur la performance du système global est illustrée par la figure 3.6. Le nombre de 4096 groupes a donné les meilleurs résultats, donc toutes les agrégations par modalité ou pour le joint ont été calculées en utilisant une représentation en sacs-de-mots avec un dictionnaire de taille 4096.

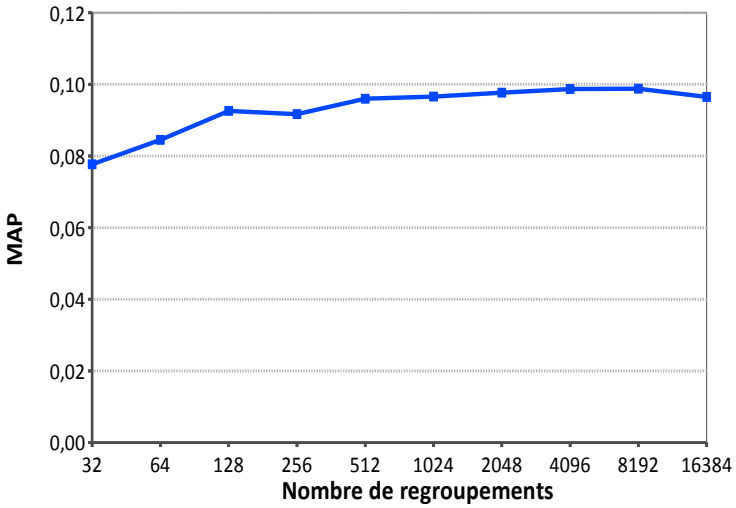


FIGURE 3.6 – La performance des MFCC en fonction du nombre de groupes.

Enfin, nous avons fixé le nombre de combinaisons des vecteurs audio-visuels considéré (n_{max}) expérimentalement par validation croisée sur l'ensemble d'apprentissage. Nous l'avons évalué avec différentes valeurs de n_{max}, pour des raisons de complexité et de temps de calcul nous nous sommes arrêtés à 32 768 combinaisons surtout que la courbe obtenue commençait à se stabiliser et que pour un n_{max} égal à 64 000 la quantité de vecteurs devenait trop importante pour un gain presque nul. Le meilleur résultat a été obtenu avec 32 768 comme le montre la figure 3.7.

3.4.3 Résultats et analyse

La métrique officielle pour cette tâche est la précision moyenne sur 100 (AP@100). Nous avons comparé la performance de différents descripteurs audio et visuels : en

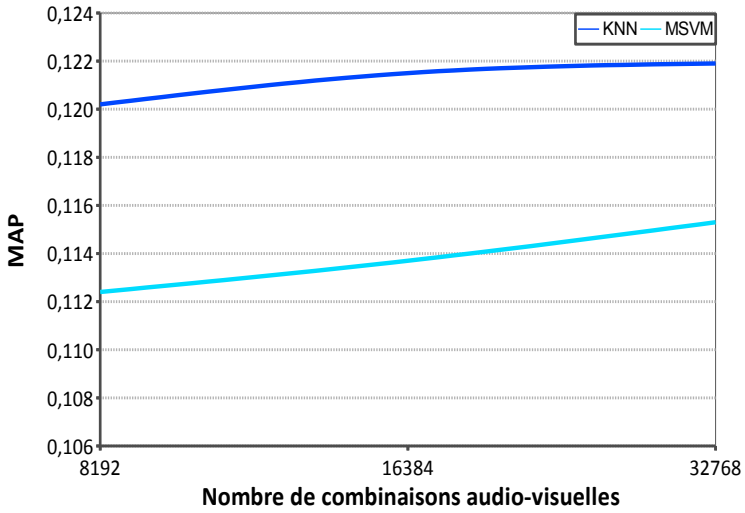


FIGURE 3.7 – L'influence du n_{max} sur la performance du descripteur audio-visuel joint.

premier temps, nous avons évalué la performance de chaque descripteur séparément. Ensuite, nous les avons opposés à la performance obtenue avec une fusion tardive en effectuant une moyenne des scores obtenus avec chaque modèle entraîné indépendamment sur chaque descripteur séparément. Enfin, nous les comparons à celles obtenues avec le descripteur joint MFCC-HOF proposé et avec la fusion du descripteur joint audio-visuel avec les deux descripteurs originaux (sacs-de-mots de MFCC et sacs-de-mots d'HOF). Nous avons rapporté dans la figure 3.8 la valeur d'AP@100 pour la détection de scènes contenant de la violence objective dans 6 des 7 films de l'ensemble de test[1]. Les résultats ont montré que le descripteur audio-visuel joint est plus performant que les deux descripteurs audio (MFCC) et visuels (HOF) séparément. Le descripteur audio-visuel joint et la fusion tardive MFCC-HOF ont obtenu globalement des résultats comparables, chacun parvenant à se démarquer sur différents films. Comme supposé, le descripteur joint audio-visuel a dépassé les différents descripteurs visuel/audio pour les films contenant une vrai cohérence entre le contenu de l'image et le signal audio comme pour *Fantastic Four1* et *Forrest Gump*. Une cohérence qui se concrétise par l'apparition d'un événement visuel en même temps qu'un événement sonore et qui a été supposée sans être quantifiée dans les films. Une mesure quantitative de cette cohérence, possiblement par l'intermédiaire d'une Analyse Canonique des Corrélation (ACC).

Pour notre soumission officielle à la campagne d'évaluation MediaEval 2013 à la tâche de détection des scènes violentes, nous avons ajouté deux autres descripteurs [Derb 13]. Le premier (OppSIFT) est basé sur le descripteur SIFT mis en place

[1]Etant donné que le septième film (Legally blonde) ne contient aucune scène violente, nous ne l'avons pas considéré.

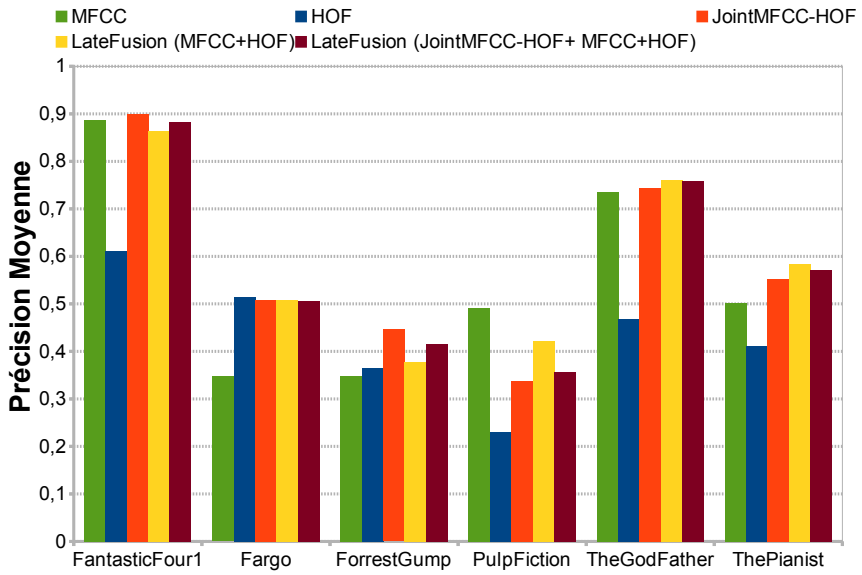


FIGURE 3.8 – L'AP@100 obtenu avec les différentes représentations en sacs-de-mots et leur fusion tardive pour la détection de scènes de violence objective sur les films de l'ensemble de test de MediaEval 2013.

par Koen van de Sande [Van 10]. Le second est un descripteur de couleur et texture (hg104) basé sur un histogramme RGB et une transformation Gabor. Dans le tableau 3.2, nous avons rapporté la précision moyenne à 100 (MAP@100) obtenu par notre soumission officielle, par la meilleure soumission et par la soumission médiane. La MAP@100 est la moyenne des AP@100 obtenus sur chaque film de la collection de test. Notre soumission a inclus la fusion des descripteurs MFCC, HOF, OppSIFT et hg104 avec le descripteur audio-visuel joint (Soumission avec jointAV). Cette soumission nous a classé premier sur 5 équipes participantes pour la détection de violence subjective (69%) et deuxième sur 9 équipes participantes pour la détection de violence objective (52%). En moyenne sur la violence objective et subjective, notre système a été capable de détecter les scènes violentes à 60.5%.

	Objective	Subjective	Moyenne
Meilleure Soumission	0.550	0.690	0.620
Soumission avec jointAV	0.520	0.690	0.605
Soumission Médiane	0.400	0.570	0.485

TABLE 3.2 – MAP@100 obtenue avec notre système à MediaEval 2013 pour la tâche de détection de scènes violentes en comparaison avec la meilleure soumission et la soumission médiane.

Les figures 3.9 et 3.10 sont fournies par les organisateurs de la tâche de détection des scènes violentes de MediaEval [Dema 14], elles comparent la performance globale des différents systèmes participants à la tâche de détection de violenceMediaE-

val 2013. Les figures 3.9 et 3.10 tracent les courbes des détections erronées/fausses alertes et du rappel/précision de la détection de la violence objective et subjective respectivement, par les système participants. Nous pouvons remarquer que notre système « LIG-run » est parmi les systèmes qui donnent le moins de detections erronées/fausse alertes et le meilleur rappel/précision pour la detections des scènes violentes objectives et subjectives.

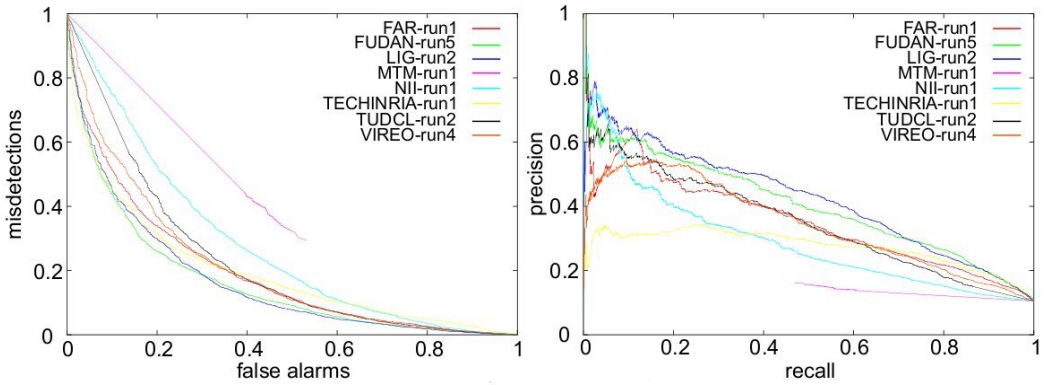


FIGURE 3.9 – Les courbes des détections erronées/fausses alertes et du rappel/précision des différents systèmes participants à la détection de la violence objective.

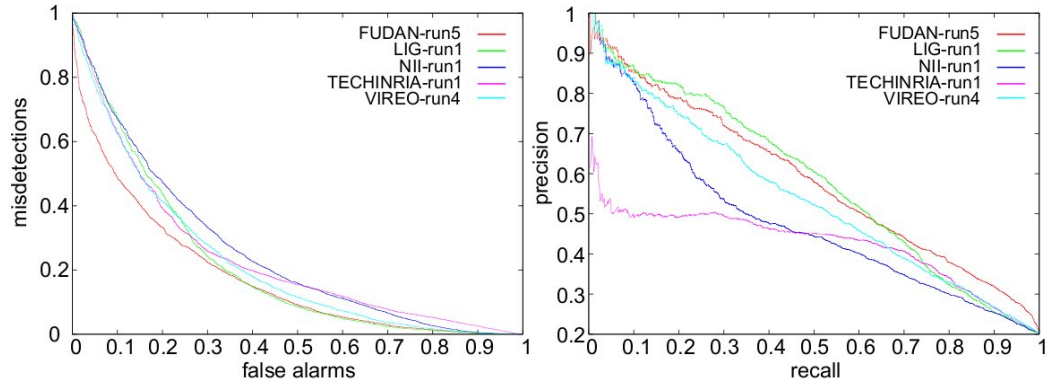


FIGURE 3.10 – Les courbes des détections erronées/fausses alertes et du rappel/précision des différents systèmes participants à la détection de la violence subjective.

Plus généralement, les courbes des détections erronées/fausses alertes montrent que la tendance est la même pour la violence objective et la violence subjective où les

meilleurs systèmes atteignent environ 20% de fausses alertes pour 20% de détections erronées pour la violence objective et environ 25% de fausses alertes pour 25% de détections erronées pour la violence subjective. En ce qui concerne le rapport rappel/précision les courbes montrent que les grandes valeurs du rappel sont atteintes au dépit de très basses valeurs de la précision (entre 0,1 ou 0,2).

3.5 Conclusion

Une nouvelle méthode a été proposée pour représenter conjointement le contenu audio-visuel dans le contexte de la détection automatique de scènes violentes. Elle exploite la corrélation entre l'information audio et l'information visuelle en construisant un dictionnaire audio-visuel joint dans le but de découvrir des motifs spécifiques audio-visuels. En comparaison avec les autres méthodes de fusion, cette méthode peut être considérée comme une fusion « doublement précoce » comme cette fusion est effectuée avant l'étape d'agrégation. Les méthodes de fusion précoces classiques, quant à elles, s'effectuent après l'étape d'agrégation et avant l'étape de classification.

La validation expérimentale sur de vrais films hollywoodiens (collection de données de MediaEval 2013) a montré que la fusion audio-visuelle jointe donne des résultats comparables à ceux obtenus avec la fusion tardive. Plusieurs pistes pourront être considérées dans le futur, comme la quantification de la cohérence entre l'image et le son dans les films et l'intégration de cette mesure dans notre système afin d'améliorer la découverte des motifs audio-visuels, l'utilisation de plus que deux descripteurs originaux (MFCC et STIP-HOF) ou encore l'application de cette représentation conjointe à d'autres types de concepts dynamiques (autres que la violence).

4

Localisations de concepts dans les images

Dans ce chapitre, nous présentons un travail réalisé sur la localisation des concepts détectés dans les images et motivé par la nouvelle sous-tâche de localisation proposé en 2013 par TRECVid dans le cadre de la tâche Semantic Indexing (SIN). Pour cela, nous proposons une approche simple et faiblement supervisée. Notre approche consiste à créer un nouveau modèle discriminant, pour un objet donné, basé sur les statistiques d'occurrence des caractéristiques locales invariantes dans les images partiellement annotées de l'ensemble d'entrainement. L'avantage principal de notre approche est son besoin limité en terme d'annotations, les images sont annotées au niveau global sans la spécification de la localisation précise des objets dans les images. Nous avons montré que notre méthode est applicable sur différents objets dans des images à conditions réelles avec des variations de positions, des arrières-plans chargés et des changements d'éclairage.
Nous avons évalué la performance de l'approche proposée sur la collection de données complexes de TRECVid 2013, et comparé notre système aux systèmes participants dans le cadre de la sous-tâche de localisation de SIN.

4.1 Motivations : Localisation des objets dans les vidéos en utilisant le moins d'annotations manuelles possible

Nous nous sommes intéressé à la localisation d'objets dans les vidéos grâce à la nouvelle sous-tâche de localisation proposé en 2013 par TRECVid dans le cadre de la tâche Semantic Indexing (SIN). C'est un défi lancé pour les systèmes participants à la tâche SIN de rendre leur détection d'objets plus précise dans le temps et dans l'espace. Dans la sous-tâche de localisation, les systèmes qui classifient les plans vidéos sont invités à déterminer la présence de l'objet dans les images-clés des plans : temporellement dans les plans *i.e.* par rapport aux images constituant le plan ; et spatialement avec un cadre englobant dans les images contenant l'objet.
En effet, la localisation des objets dans les images est un problème très important dans le domaine de traitement d'images. Elle est indispensable pour différentes applications de l'analyse automatique d'images comme la séparation des objets de l'arrière-plan ou l'extraction des relations spatiales entre différents objets dans une même image. Les systèmes de détection d'objets sont censés trouver les images contenant des instances de la classe ou la catégorie d'objets considérée et donner leur emplacement dans ces images. Par conséquent, la résolution du problème de localisation nécessite un système non seulement capable de reconnaître les objets mais aussi d'indiquer leur emplacement précis dans les images par l'intermédiaire d'un cadre englobant, d'un polygone ou bien un masque de pixels. La localisation d'objets est une tâche très difficile à cause de la variabilité du contenu des images : des changements de point de vue, des déformations des objets, des changements d'illumination, ou encore de l'occlusion.

Les approches existantes résolvent le problème de localisation selon différents

niveaux de supervision. Nous pouvons distinguer entre les techniques de localisation d'objets fortement supervisées (très communes) et celles qui sont faiblement supervisées. Les techniques fortement supervisées exigent un ensemble d'images positives et négatives annotées et avec un marquage précis de l'emplacement des objets pour la phase d'apprentissage. Les techniques faiblement supervisées, quant à elles, visent à effectuer la même tâche mais sans aucun marquage de l'emplacement des objets. Comme les objets peuvent apparaître plusieurs fois et de façon arbitraire dans les images positives avec un arrière-plan encombré, la tâche de localisation devient plus facile avec un niveau élevé de supervision.

La solution la plus simple pour la localisation d'objets, serait donc d'annoter et de marquer manuellement les objets de toutes les données d'entraînement (un apprentissage fortement supervisé). Récemment, des travaux approfondis sur des corpus de tailles limitées, comme Caltech04 [Zhan 10, Nguy 09, Opel 05] ou Weizmann [Winn 05b] ou juste quelques catégories PASCAL-VOC [Zhan 10, Pand 11], ont montré que l'apprentissage supervisé est une approche très prometteuse pour résoudre la localisation d'objets. Bien que cette approche soit efficace pour les petit corpus ou pour les objets qui sont bien définis et facilement délimitables, elle n'est pas générique ni applicable sur les grands corpus (comme TRECVid). Par conséquent, l'apprentissage faiblement supervisé s'impose comme sa plus grande alternative pour réduire les coûts [Gall 08, Ferg 07, Zhan 10, Pres 12, Cran 06, Nguy 09, Opel 05].

Nous nous sommes intéressés à la question suivante : dans quelle mesure les méthodes faiblement supervisées peuvent-elles contribuer à la localisation d'objets dans les grandes collections de vidéos ? Afin de tenter de répondre à cette question nous procédons de la façon suivante :

- Nous commençons par proposer un cadre applicatif basé sur des techniques d'apprentissage faiblement supervisé pour la détection et la localisation d'objets dans les vidéos réelles.
- Ensuite, nous évaluons la méthode proposée sur le jeu de données TRECVID 2013 et nous comparons nos performances à celles des approches hautement supervisés.

4.2 Travaux connexes

Plusieurs méthodes basées sur des techniques de segmentation ont été proposées pour la localisation d'objets [Fuss 06, Russ 06, Todo 06, Winn 05b]. Ces méthodes offrent une bonne localisation de l'objet en déterminant le contour des objets et en général elles sont en mesure de traiter le problème de déformation des objets dans les images. La faiblesse de ces méthodes vient du besoin de paramétrer à l'avance la forme des objets à segmenter parce qu'elles ne peuvent pas traiter directement toutes les formes possibles. Liebe *et al.* ont construit un dictionnaire pour chaque catégorie d'objets à

l'aide d'une représentation basée sur les différentes parties de l'objet (ou part-based representations). Ils ont utilisé ensuite un modèle de forme implicite pour segmenter automatiquement les objets [Leib 04].

L'approche la plus commune pour la localisation d'objets reste l'approche de fenêtre glissante. Dans cette approche, une fenêtre est glissée le long de l'image à différentes échelles et un classificateur est appliqué à chacune des sous-images obtenues par les fenêtres glissantes ; le maximum des scores de classification est considéré comme une indication de la probabilité de présence de l'objet dans la région de l'image [Harz 09, Lamp 08, Bosc 07, Chum 07, Ferr 08, Rowl 95]. Les méthodes basées sur les parties déformables (ou Deformable Part-Based Models-DPM) constituent les méthodes de pointe pour la détection d'objets par le glissement d'une fenêtre. Une DPM représente un objet par un filtre initial de basse résolution disposé dans une configuration spatiale flexible. De nombreux travaux récents utilisent cette technique [Amit 07, Bern 05, Rama 06]. Felzenszwalb *et al.* ont montré les bons résultats de cette technique pour la localisation d'objets sur les données PASCAL VOC [Felz 10]. L'inconvénient des approches à fenêtres glissantes est leur coût en temps de calcul. Lampert *et al.* ont proposé une méthode efficace pour la localisation et la détection simultanée d'objets en réduisant le nombre de fenêtres qui doivent être traitées pour accélérer la localisation des objets [Lamp 08]. La performance de localisation de cette approche est bonne, mais l'approche reste fortement supervisée.

Les approches faiblement supervisées suscitent de plus en plus d'intérêt dans ce domaine. L'une des raisons est le besoin de réduction des travaux d'annotations humains. D'un certain point de vue, notre problème peut être considéré comme un problème d'Apprentissage d'Instances Multiples (MIL). Cette catégorie de problème d'apprentissage découle des cas d'applications où les exemples d'entraînement sont ambigus ou pas entièrement annotés [Diet 97]. Au lieu d'analyser des ensembles d'exemples annotés comme positifs (tous les exemples contenus sont positifs) et négatifs (tous les exemples contenus sont négatifs), la méthode traite des « sacs d'exemples » positifs (au moins un exemple du sac est positif) ou négatifs (tous les exemples sont négatifs). Ensuite, une méthode spécifique est utilisée pour trouver les éléments communs entre les exemples positifs qui ne figurent pas dans les exemples négatifs. Comme la localisation d'objets faiblement supervisé ne nécessite que des images annotées au niveau de l'image pour l'apprentissage, les exemples d'apprentissage positifs contiennent en réalité des exemples positifs et négatifs.

Actuellement, la grande majorité des techniques de localisation faiblement supervisées ont été appliquées à des données relativement faciles. Ceci est dû à la démultiplication des principaux problèmes de localisation d'objets (comme les changements de positions, d'échelles ou de poses) qui limitent son efficacité. Ries *et al.* ont proposé une méthode faiblement supervisée pour localiser des logos et des fleurs dans les images. Ils ont créé un modèle de couleur discriminant pour un objet donné à partir de statistiques d'occurrence de ses couleurs dans les images [Ries 12]. Leur

méthode est très intéressante car les modèles de couleur discriminants sont relativement rapides à calculer et donc pratiques pour la localisation de concepts dans les grands corpus d'images. Ils permettent d'éliminer rapidement les images négatives ambigües en amont du processus de classification. Par contre il n'est pas adapté pour les images à condition réelles présentant des occlusions et des arrière-plans chargés et où les objets considérés apparaissent dans un grand nombre de couleurs différentes dans des environnement très variés.

4.3 Modèle de détection et de localisation des objets dans les vidéos

Notre but est de localiser des instances d'objets (ou concepts) dans les vidéos indépendamment des variations d'échelles, de point de vue, d'orientation et d'illumination dans des données complexes comportant des occlusions avec des arrière-plans chargés. Comme défini dans la sous-tâche de localisation (6.1), nous effectuons la localisation dans les images-clés des plans mais nous décidons de le faire d'une façon faiblement supervisée, c'est-à-dire en utilisant uniquement des annotations fournies au niveau des images sans aucun marquage de l'emplacement précis des objets pour la phase d'apprentissage.

Notre approche tente de détecter de l'invariabilité spécifique à un concept dans la variabilité globale. Elle consiste à construire un modèle discriminant en utilisant les statistiques d'occurrence des caractéristiques locales invariantes. Il s'agit d'une approche faiblement supervisée en deux étapes. Dans un premier temps, pour chaque classe d'objet, nous retrouvons les images contenant l'objet cible. Dans un deuxième temps, nous localisons les objets dans ces images. Cette approche est illustrée dans la figure 4.1.

FIGURE 4.1 – Le système global de localisation d'objet.

4.3.1 Système de détection d'objets

Le système de détection d'objets employé respecte l'architecture du processus général détaillé dans 2.3. Le but de l'utilisation du système de détection d'objet est de retrouver les images contenantes l'objet cible et donc les images sur lesquelles la localisation doit être effectuée.

4.3.2 Système de localisation d'objets

Le système de détection d'objets est capable d'apprendre et de classifier les images. Néanmoins, la classification obtenue manque d'information plus précise concernant la localisation de l'objet dans l'image. En effet, les objets peuvent être situés n'importe où dans les images sélectionnées. Quelle partie de l'image représente l'objet désiré ? Où est l'objet désiré et où est l'arrière-plan ? Pour répondre à ces questions, nous développons un modèle pour la localisation d'objets basé sur la statistique d'occurrence de descripteurs locaux.

Notre contribution principale concerne la création d'un nouveau modèle discriminant pour la localisation d'objets basé sur la statistique d'occurrence des caractéristiques locales invariantes à partir d'un ensemble d'images positives (contenant l'objet considéré) et négatives (ne contenant pas l'objet considéré). L'idée de cette méthode est proche de celles qui représentent des objets à partir de leurs informations de couleur et attribue une valeur de classement (0 ou 1) pour chaque valeur de couleur de l'espace de couleur considéré. Ries *et al.* ont calculé un modèle de couleur à l'aide des statistiques d'occurrence des couleurs des images annotées d'apprentissage [Ries 12]. Cependant, leur approche repose sur deux hypothèses. Premièrement, les objets considérés doivent apparaître dans un nombre limité de couleurs. Deuxièmement, les couleurs de l'arrière-plan des images positives et négatives doivent être similaires et plus variés que celles de l'objet recherché. Ces hypothèses constituent la principale limitation de leur modèle, car elles ne sont manifestement pas vérifiées dans les images réelles où les objets apparaissent en différentes positions dans des environnements complexes. Pour ces raisons, nous avons choisi de représenter le contenu de l'image avec un descripteur invariant local pour notre modèle discriminant.

L'approche de localisation d'objets que nous avons proposé s'effectue en trois étapes. Nous représentons, d'abord, le contenu de l'image selon des descripteurs locaux invariants. Nous créons, ensuite, le modèle discriminant à partir de la fréquence d'apparition des descripteurs calculés pour identifier ceux qui sont distinctifs pour un objet spécifique. Enfin, nous cherchons le meilleur cadre englobant l'objet dans les images à l'aide de l'occurrence des descripteurs distinctifs.

L'extraction des cadres englobants est faite séparément pour chaque objet (ou concept), donc la méthode est décrite ci-dessous pour un concept donné et sur l'ensemble des images annotées I avec I_p l'ensemble des images positives et I_n l'ensemble des images négatives de l'ensemble d'apprentissage.

4.3.2.1 Extraction de descripteurs locaux

Comme mentionné précédemment, dans les images réelles les objets ciblés apparaissent dans un grand nombre de couleurs différentes et dans des environnements aussi variés que chargés. Pour surmonter les problèmes de diversification des cou-

leurs et être plus robuste aux changements de taille et de point de vue, nous avons choisi d'utiliser une représentation locale.

Pour la sélection des régions saillantes (points d'intérêt), nous utilisons un détecteur invariant à l'échelle, par exemple le détecteur de Harris-Laplace. Ensuite, nous représentons chaque point d'intérêt par l'intermédiaire d'un descripteur, par exemple le Opponent Scale Invariant Feature Transform (SIFT) [Van 10].

4.3.2.2 Création du modèle discriminant

L'idée principale est de déterminer les SIFTs discriminants pour un objet donné, en d'autres termes trouver les SIFTs qui apparaissent significativement plus souvent dans des images positives que dans des images négatives. Le modèle discriminant est calculé à partir des statistiques d'occurrence des SIFTs dans un ensemble I_p d'images positives et un ensemble I_n d'images négatives.

Une fois que les points d'intérêt pour chaque image sont calculés et représentés par le descripteur Opponent SIFT, nous appliquons une méthode de regroupement (ou clustering) standard sur tous les points SIFTs de toutes les images des données d'entraînement en différents groupes (ou clusters). Il en résulte l'attribution d'un cluster spécifique c à chaque point SIFT, avec $c \in C$ et C est l'ensemble des groupes. L'ensemble de ces groupes formeront les mots du dictionnaire visuel. Pour identifier les mots visuels les plus représentatifs pour un objet donné, nous associons à chaque mot visuel c ses fréquences d'occurrence relatives (ROF). Les ROFs représentent le pouvoir discriminant du mot visuel c, c'est-à-dire à quel point un mot visuel c est pertinent pour un objet. Souvent, plus le mot visuel c apparaît dans les images positives, plus il est susceptible d'indiquer l'objet en question. Par conséquent, pour un objet donné, nous calculons les fréquences d'occurrence relative (ROF) avec laquelle chaque groupe de SIFT est présent dans les images positives et négatives.

La fréquence d'occurrence relative (ROF) de chaque mot visuel (SIFT cluster) c est donnée respectivement pour I_p et I_n par :

$$ROF_p(c) = \frac{|\{I \in I_p | y \in f(S(I))\}|}{|I_p|} \tag{4.1}$$

$$ROF_n(c) = \frac{|\{I \in I_n | y \in f(S(I))\}|}{|I_n|} \tag{4.2}$$

$$f(S(I)) = \{c \mid \exists s \in S(I) \; et \; s \in C\} \tag{4.3}$$

Où $S(I)$ est l'ensemble des points SIFT de l'image I, f est la fonction qui permet de retourner le cluster c auquel un point SIFT s appartient et $f(S(I))$ est la projection de $S(I)$ sur C.

La formule (4.1), respectivement (4.2), représente le quotient du nombre absolu d'images positives (resp. négatives) dans laquelle au moins un point d'intérêt appartenant au mot visuel c est présent sur le nombre total d'images positives (resp. négatives), dans l'ensemble des données d'entraînement.

Pour identifier le mot visuel le plus discriminant, nous comparons son ROF en images positives et négatives. Si le mot visuel c apparaît plus fréquemment dans les images positives que dans les images négatives ($ROF_p(c) <= ROF_n(c)$), cela signifie que c est un mot visuel plus spécifique que le reste des mots visuels et représentatif pour l'objet considéré. Nous pouvons représenter la relation entre $ROF_p(c)$ et $ROF_n(c)$ par la formule suivante :

$$(ROF_p(y)/ROF_n(y)) > 1 \tag{4.4}$$

La formule (4.4) pourrait être un excellent indicateur du pouvoir discriminatoire du mot visuel c pour des données d'entraînement équilibrées c'est-à-dire où la quantité d'images positives et celle des images négatives sont approximativement équivalentes. Cependant, la plupart du temps la quantité d' images négatives est sensiblement plus élevée que celle des images positives. Par conséquent, nous avons décidé de ne considérer que la ROF_p pour chaque c.

4.3.2.3 La recherche du meilleur cadre englobant

Pour localiser un objet donné dans une image, nous utilisons les cadres englobants rectangulaires traditionnels. Nous détaillons ci-dessous la méthode que nous proposons pour la recherche du meilleur cadre englobant l'objet ciblé.

Évidemment, pour une localisation avec un cadre englobant rectangulaire, nous devons déterminer deux points du rectangle autour de l'objet à localiser dans une image. Idéalement, la bonne localisation est obtenue lorsque les points SIFTs se trouvent tous sur l'objet dans l'image. Il est alors facile de déterminer le rectangle qui couvre ces SIFTs. Cependant, en pratique, les images contiennent de nombreux objets et les points SIFTs sont éparpillés sur toute l'image. Ainsi, il est préférable de filtrer les points SIFTs de chaque image de façon à ne garder que les SIFTs intéressants : ceux qui se trouvent sur l'objet ciblé.

Nous décrivons ici la méthode de filtrage proposée pour retrouver les extrémités gauche et droite du cadre englobant. Tout d'abord, les points SIFTs sont filtrés pour conserver seulement ceux qui sont associés à un groupe c dont le ROF_p est supérieur à un seuil minimum α. F(I) représente l'ensemble des points SIFTs conservés et se calcule de la manière suivante :

$$F(I) = \{s \in S(I) | ROF_p(f(s)) > \alpha\} \tag{4.5}$$

Ensuite, nous découpons la largeur de l'image (w) en K intervalles égales (T_k) et nous calculons l'histogramme de la projection sur l'axe horizontale des points filtrés

sur les K différents intervalles.

$$T_k = [\frac{k}{K}w, \frac{k+1}{K}w[\ ,\ 0 \leq k \leq K-1 \tag{4.6}$$

L'histogramme est calculé selon la formule suivante :

$$h(k) = \frac{\sum s \in F(I) \mid x(s) \in T_k}{|F(I)|} \tag{4.7}$$

Enfin, les extrémités gauche et droite du rectangle sont déterminées en partant du bord et en le décalant tant que les cases de l'histogramme ne sont pas suffisamment remplies et donc tant que la valeur de la case est inférieure au seuil β ; dans le cas où elles sont toutes inférieures à β le cadre englobant sera éliminé. S'il y a au moins une case supérieure à β, le rectangle qui couvre les cases restantes est finalement conservé et considéré comme le cadre englobant de l'image-clé. Pour une meilleure localisation le seuil β doit être fixé pour chaque objet séparément. Les extrémités droites et gauches du rectangle sont donc fixées comme suit :

$$k_{left} = \max_{k \in [0,K-1]} \{k \mid \forall\, l < k, h(l) < \beta\} \tag{4.8}$$

$$k_{right} = \max_{k \in [0,K-1]} \{k \mid \forall\, l > k, h(l) < \beta\} \tag{4.9}$$

Les extrémités haute et basse du cadre englobant sont fixées de la même manière en calculant l'histogramme de la projection sur l'axe verticale. La méthode est illustrée dans la figure 4.2.

4.4 Évaluation du système proposé

La figure 4.2 illustre un exemple de l'algorithme de filtrage et le cadre englobant que nous proposons. Comme la figure montre, il y a beaucoup de points SIFTs (les points en bleu), le cadre rectangulaire externe (en vert) présente le résultat de la localisation de l'objet *Motorcycle* en prenant tous les points SIFT de l'image. Le cadre rectangulaire interne (en bleu) est la localisation obtenue après le filtrage que nous proposons. Notre système de localisation est alors bien capable de localiser l'objet dans l'image mais ne parvient pas à dessiner le cadre englobant l'intégralité de l'objet. Comme nous pouvons le voir, le filtrage contribue à fournir une localisation plus précise des objets, en particulier, lorsque l'objet apparaît comme objet principal de l'image.

Afin d'évaluer l'efficacité et la précision d'un système faiblement supervisé pour la détection et la localisation des objets dans les vidéos, nous avons choisi une grande collection de vidéos complexes avec un grand nombre d'objets et des arrière-plan chargés : TRECVid 2013. Nous avons effectué cette évaluation en deux temps : premièrement en comparant les résultats obtenus par notre système à ceux obtenus par une méthode de localisation de référence qui considère le centre des images,

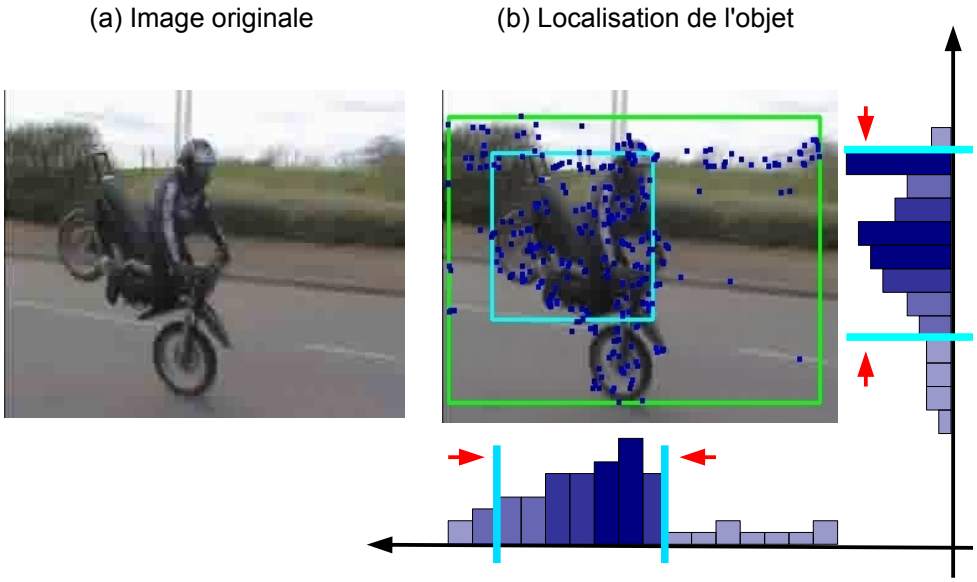


FIGURE 4.2 – Exemple de la localisation de l'objet *Motorcycle* en utilisant les ROF : (a) est l'image originale ; (b) l'image avec les points SIFTs correspondants et deux cadres rectangulaires englobants autour de l'objet, le cadre externe (en vert) est le cadre de référence obtenu en gardant tous les points SFITs (baseline) alors que le cadre interne (en bleu) est obtenu avec notre méthode avec un seuil $\beta = 0.25$.

deuxièmement en comparant la performance de notre système à celles des systèmes participants à la sous-tâche de localisation de TRECVid 2013.

La suite de la section est organisée de la façon suivante, nous commençons par présenter la collection de données TRECVid 2013 et plus précisément la tâche à laquelle nous avons participé. Nous abordons ensuite les caractéristiques nécessaires aux systèmes participant ainsi que les métriques officielles. Enfin, nous précisons les choix des paramètres de notre système et nous exposerons les résultats obtenus par notre système et nous proposons une comparaison avec les systèmes fortement supervisés participants à la même tâche.

4.4.1 TRECVid 2013

Pour évaluer la qualité de la méthode proposée, nous avons choisi la nouvelle sous-tâche de localisation de la tâche Semantic Indexing (SIN) de la campagne d'évaluation TRECVID 2013. La localisation est limitée à dix objets choisis parmi ceux de la tâche principale SIN. Ces objets sont : *airplane, soat_ship, bridge, bus, chair, hand, motorcycle, telephones, flags* et *quadruped*. Nous avons évalué notre approche sur ces dix classes d'objets et comparé nos résultats à ceux des systèmes participants.

La collection de données TRECVID 2013 se compose de deux grands sous-ensembles. L'ensemble d'entraînement qui contient 545 923 plans de vidéos annotés et l'ensemble de test qui contient 112 677 plans de vidéos. Les plans de l'ensemble d'entraînement sont annotés comme contenant l'objet ou non, sans aucune information concernant le lieu exact d'apparition de l'objet dans les images du plan.

4.4.2 Système

La tâche de localisation nécessite un système capable de trouver les coordonnées (x, y) des deux points définissant un rectangle autour de l'objet cible [Over 13]. Ceci doit être fait pour chaque classe d'objets spécifiée dans la liste et pour chaque image clé (dite I-frame) des 1000 premiers plans classés par le système de détection d'objets. Le cadre englobant doit inclure l'objet en entier et être en même temps le plus petit possible. Dans le cas où il y a plusieurs instances de l'objet dans l'image, le système peut proposer plusieurs cadres englobants mais un seul sera pris en compte pour l'évaluation finale, sachant que la vérité terrain fournie par les organisateurs TRECVID ne propose qu'un seul cadre par image. Notre modèle ne traite pas ce cas de figure pour le moment, il propose un seul cadre englobant par image.

4.4.3 Métriques d'évaluation

Dans l'ensemble de données de test, pour chaque plan de vidéos positif (jugé comme contenant l'objet) un sous-ensemble d'images du plan (I-frames) sont consultées et annotées manuellement pour localiser les pixels représentant l'objet. L'ensemble d'images annotées sont ensuite utilisées pour évaluer la localisation fournie par les

systèmes participants [Over 13]. La qualité de la localisation est évaluée temporellement et spatialement, à l'aide de la précision et du rappel à deux niveaux, celui de l'image et celui du pixel respectivement. Ainsi, deux métriques différentes sont calculées « I-frame Fscore » et « mean pixel Fscore ». « I-frame Fscore » représente la qualité de la localisation temporelle et évalue la capacité du système de détection d'objet à retrouver les plans et les images des plans (I-frame) contenant l'objet. Les métriques telles qu'elles sont définies par TRECVID sont calculées selon les formules suivantes où TP est le nombre d'éléments correctement attribués à une classe d'objet par le système évalué, n est le nombre total d'éléments retournés pour un objet donné et BB_pixel sont les pixels du cadre englobant :

$$\text{rappel I-frame} = \frac{\#\text{TP I-frames retournés}}{\#\text{total I-frames pertinents}} \tag{4.10}$$

$$\text{précision I-frame} = \frac{\#\text{TP I-frames retournés}}{\#\text{total I-frames retournés}} \tag{4.11}$$

$$\text{I-frame Fscore } = 2.\frac{\text{rappel I-frame.précision I-frame}}{\text{rappel I-frame} + \text{précision I-frame}} \tag{4.12}$$

La « mean pixel Fscore » représente la qualité de la localisation spatiale et évalue plus précisément la capacité du système de localisation à trouver le meilleur cadre englobant les objets dans les I-frame détectés. Cette métrique est calculée à partir de la moyenne du rappel niveau pixels (MPR) et la moyenne de la précision niveau pixels (MPP) :

$$\text{MPR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\#\text{TP BB\_pixel retournés}}{\#\text{total BB\_pixel pertinents}} \tag{4.13}$$

$$\text{MPP} = \frac{1}{n}\sum_{i=1}^{n}\frac{\#\text{TP BB\_pixel retournés}}{\#\text{total BB\_pixel retournés}} \tag{4.14}$$

$$\text{Mean pixel Fscore} = 2.\frac{\text{MPR.MPP}}{\text{MPR} + \text{MPP}} \tag{4.15}$$

Plus la valeur de ces deux métriques est élevée, meilleure est la qualité du système de localisation d'objets dans les plans. Une moyenne pour chacune de ces métriques est calculée pour chaque objet séparément et sur l'ensemble des objets considérés.

4.4.4 Réglages de paramètres

Pour notre système de détection, nous disposons de 15 différents types de descripteurs calculés et fournis par IRIM et XEROX [Ball 13, Hama 13]. Ces descripteurs sont ensuite optimisés par la méthode proposée par Safadi *et al*, qui combine une réduction des dimensions basée sur une PCA avec une transformation de puissance. Pour la classification, deux méthodes d'apprentissage différentes ont été utilisées, l'une basée sur plusieurs SVM pour gérer le problème de déséquilibre entre classe et une autre basée sur les K plus proches voisins. La fusion des descripteurs et la

fusion hiérarchique ont permis de combiner les scores de prédictions des différents classificateurs des différents descripteurs. Enfin une méthode de reclassement temporel (proposée par Safadi [Safa 11b]) a été appliquée pour établir une liste des images classées par ordre décroissant des scores de prédiction de présence de l'objet cible.

Pour notre système de localisation, nous appliquons un regroupement des SIFTs en 4096 groupes (C à 4096) et nous fixons la taille des histogrammes de ROF à 32 éléments. Étant donné que cette édition de la sous-tâche de localisation est la première dans TRECVID, nous n'avions aucune information concernant les métriques officielles pour l'évaluation des systèmes participants lors de la mise en place de notre modèle. Par conséquent, nous fixons les paramètres de notre modèle manuellement en vérifiant certains cadres englobant fournis par notre système. Le paramètre β de la méthode de localisation a été réglé et optimisé sur l'ensemble d'apprentissage, en examinant visuellement la localisation dans les 500 premières images retournées par notre système pour chaque objet.

4.4.5 Résultats et analyse

Dans cette section, nous présentons la performance de notre système de localisation sur la collection de données TRECVID 2013 pour les dix concepts (ou objets) spécifiés. Nous avons appliqué l'algorithme de localisation sur chaque image clé (I-frame) des 1000 premiers plans retournés par notre système de détection de concepts (SIN).

La figure 4.3 montre quelques exemples de la tête du classement des résultats obtenus pour les dix objets considérés. Nous avons constaté que le système de détection retournait les bons plans pour la plupart des objets. De plus, la localisation pour les objets comme *flags, hand* et *chair* sont meilleurs que ceux obtenus pour les objets comme *telephone* et *bus* où la localisation était beaucoup moins exacte. Ceci peut être expliqué par le fait que ces objets n'apparaissent pas comme des objets principaux dans les données d'apprentissage contrairement aux autres objets qui eux apparaissent clairement et qui sont représentés par une grande partie des pixels des images. Par exemple l'objet *telephone* apparaît dans les images en très petite taille et souvent dans la main des personnes ou avec d'autres objets bien plus imposants. Cependant, l'algorithme de localisation a montré une bonne capacité à localiser l'objet principal dans les images, même dans le cas où le système de détection retournait de mauvaises images. Par exemple, pour les objets *bus* et *boat_ship*, quelques images retournées par le système de détection sont mauvaises et ne contiennent pas ces objets mais l'algorithme de localisation réussissait à encadrer l'objet principal dans ces images.

Le tableau 4.1 compare la qualité de la localisation obtenue par une méthode de référence (baseline) et notre approche basée sur un modèle de descripteurs locaux discriminant, en termes de « mean pixel Fscore ». Cette métrique est plus pertinente pour notre travail car elle évalue, plus précisément, la qualité de la localisation des objets dans les images par les différents systèmes. Comme résultat de référence,

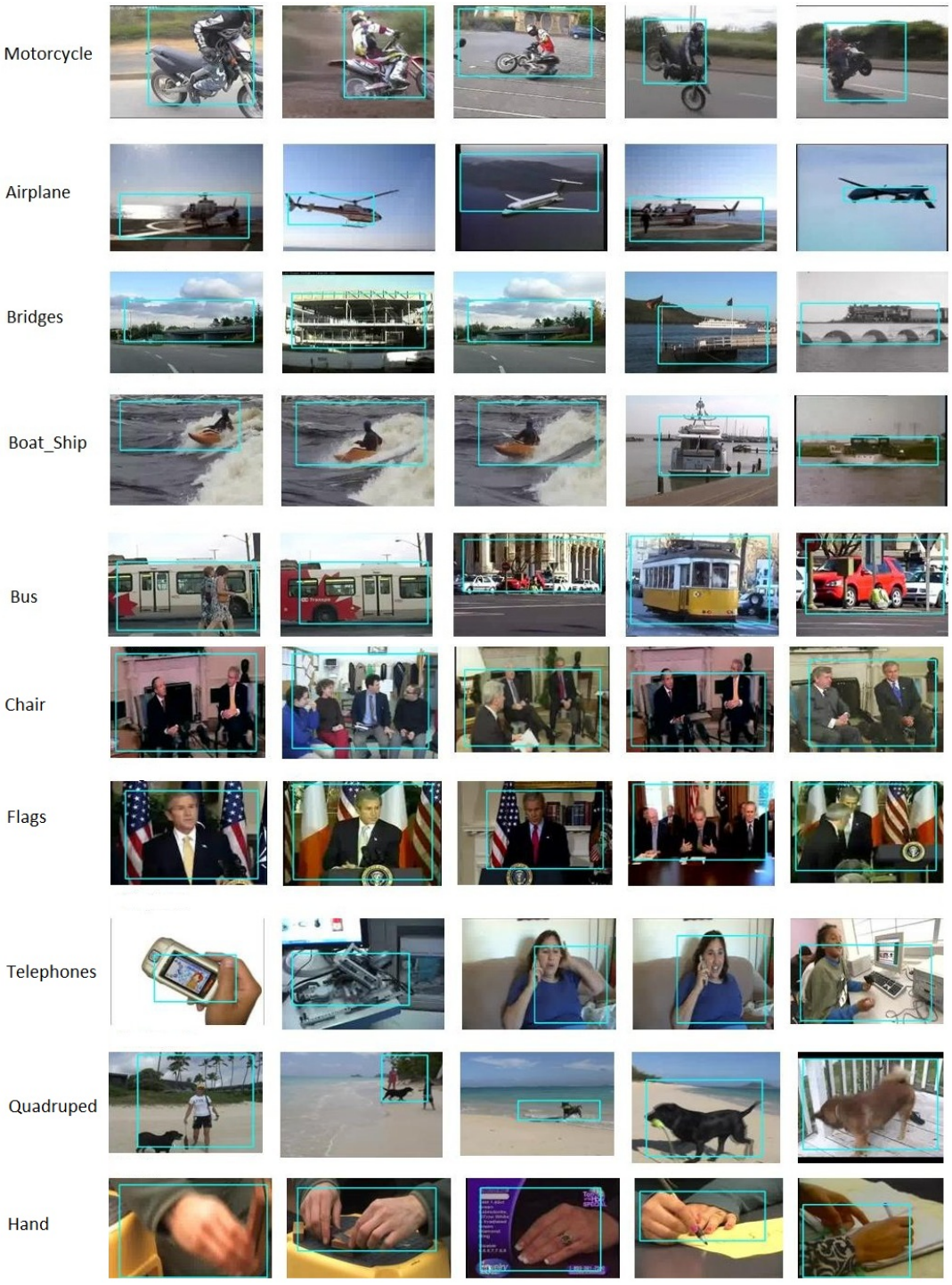


FIGURE 4.3 – Quelques exemples des résultats obtenus avec notre système global pour la détection et la localisation des objets dans les plans vidéos.

Objet	Baseline	Notre modèle
Airplane	0.39%	**0.53%**
Boat_Ship	4.45%	**5.39%**
Bridges	0.87%	**2.02%**
Bus	0.24%	**0.41%**
Chair	4.49%	**6.72%**
Hand	8.19%	**15.18%**
Motorcycle	**3.00%**	2.53%
Telephones	**0.29%**	0.24%
Flags	1.49%	**6.17%**
Quadruped	2.75%	**3.79%**
Mean	2.61%	**4.29%**

TABLE 4.1 – Comparaison de la qualité de notre système global avec une méthode basique de référence (baseline) selon la métrique de « mean pixel Fscore ».

nous utilisons celui obtenu avec un cadre englobant situé au centre de l'image. Cette méthode est généralement considérée comme un bon indicateur de l'emplacement des objets, car la plupart des objets ont tendance à se trouver au centre de l'image. Les résultats montrent qu'une amélioration significative est obtenue avec notre méthode en comparaison à l'approche de référence.

Nous avons comparé également nos résultats à ceux des équipes participantes à la même sous-tâche de localisation de la tâche SIN de TRECVID 2013. Quatre équipes ont participé : FTRDBJ représente Orange Labs à Pékin, SRI_Aurora représente l'Université Centrale de Floride (UCF), Amsterdam représente l'Université d'Amsterdam (UvA) et nous. La figure 4.4 montre la performance des systèmes participants en termes de localisation temporelle d'objets dans les plans sur les des données de test. Il contient le « I-frame Fscore » obtenu par chaque système participant pour les 10 concepts considérés (objets) et leur moyenne. Comme précisé précédemment, cette métrique évalue la qualité du système de détection d'objet. Comme nous pouvons remarquer, nous avons obtenu le deuxième meilleur résultat en termes de localisation temporelle avec une moyenne d'I-frame F-score égale à $16,51\%$ derrière l'équipe d'Amsterdam qui a obtenu les meilleurs résultats avec une moyenne égale à $23,36\%$.

La figure 4.5 montre la performance des systèmes participants en termes de localisation spatiale des objets dans les images sur les données de test. Il reprend la « Mean Pixel Fscore » obtenue par chaque équipe participante pour les 10 objets considérés et leur moyenne. Cette mesure évalue directement la qualité du système de localisation. Par rapport à cette métrique, notre système a obtenu la troisième place avec une « Mean Pixel Fscore » égale à $4,3\%$ derrière l'équipe FTRDBJ $(9,6\%)$ et l'équipe d'Amsterdam (11%). Contrairement à notre méthode faiblement supervisé,

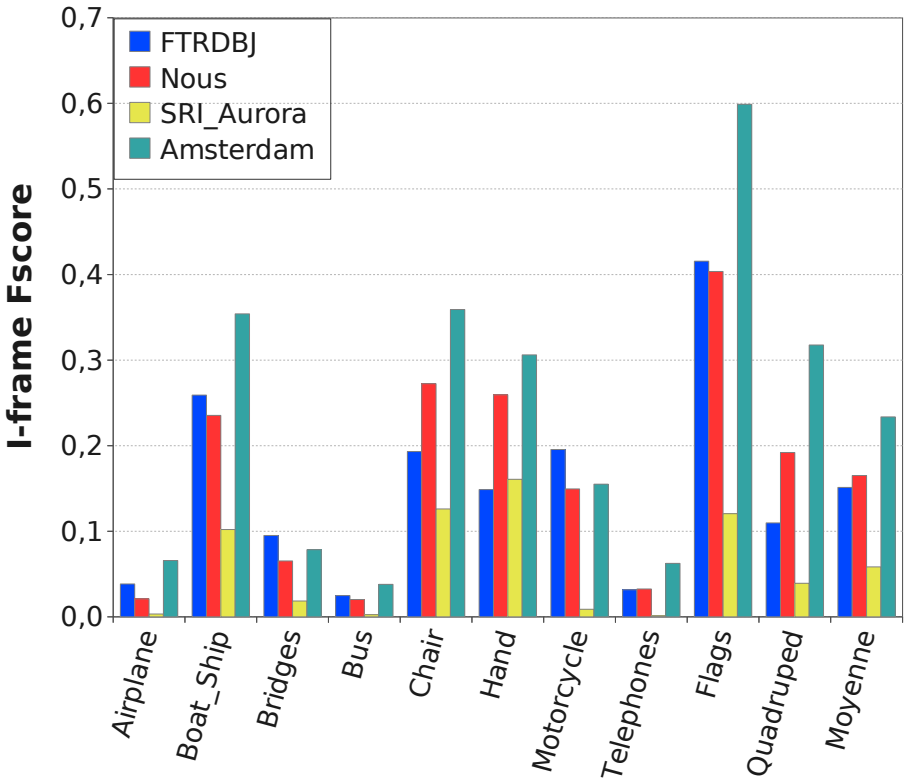


FIGURE 4.4 – Résultats officiels obtenus par les systèmes participants à la sous-tâche de localisation à TRECVid 2013, en termes de « I-frame Fscore ». Le schéma reprend les résultats par concept séparément et la moyenne sur les dix concepts considérés.

nous soulignons que tous les autres participants ont proposé des systèmes de localisation fortement supervisés en ajoutant des annotations manuelles marquant l'emplacement exact des objets dans les images d'apprentissage [Snoe 13, Dehg 13, Bai 13].

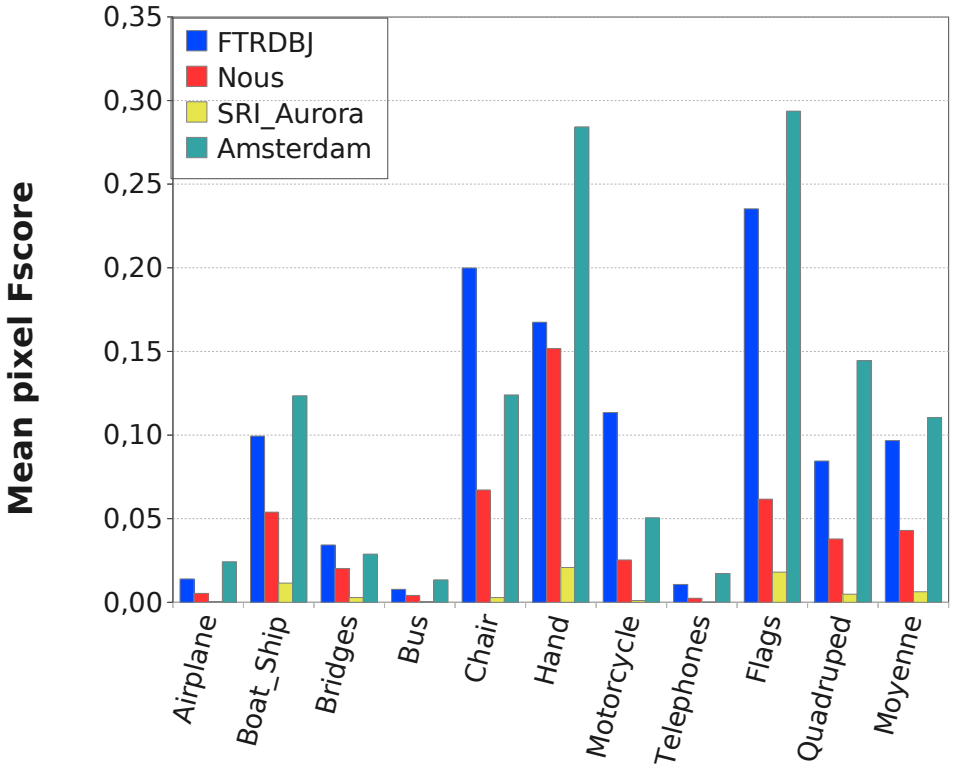


FIGURE 4.5 – Résultats officiels obtenus par les systèmes participants à la sous-tâche de localisation à TRECVid 2013, en termes de « Mean Pixel Fscore ». Le schéma reprend les résultats par concept séparément et la moyenne sur les dix concepts considérés.

Étant donné que les systèmes proposés par l'équipe FTRDBG et l'équipe d'Amsterdam sont des systèmes fortement supervisés, il est difficile d'analyser les résultats obtenus. Cependant, nous observons que pour certains objets difficiles comme *avion, bus et téléphone* même les méthodes fortement supervisées n'ont pas réussi à obtenir de meilleurs résultats que notre méthode faiblement supervisée. D'un autre côté, pour les petits objets qui apparaissent généralement comme des objets secondaires, les méthodes supervisées dépassent la méthode faiblement supervisée. La meilleure précision des méthodes fortement supervisées est la conséquence logique d'un apprentissage sur des données plus exactes et plus propres grâce aux annotations manuelles supplémentaires encadrant les objets dans les images. Néanmoins, nous pensons qu'il est intéressant de souligner que le système proposé reste prometteur pour un méthode faiblement supervisée.

4.5 Conclusion

Nous avons proposé ici une méthode faiblement supervisée pour la localisation d'objets dans les vidéos réelles. Notre méthode consiste à créer un nouveau modèle discriminant basé sur l'occurrence statistique de descripteurs locaux invariants à partir de données d'entraînement faiblement annotées (le lieu d'apparition des objets dans les images n'est pas connu). Ces images faiblement annotées sont la seule quantité de supervision requise pour notre approche. Les principaux avantages de notre système sont le faible niveau de supervision nécessaire et sa capacité à traiter des vidéos non segmentées et chargées. En outre, le modèle d'apprentissage est indépendant des variations de couleur de l'objet et l'arrière plan dans les images. Il intègre également une certaine robustesse aux changements d'échelle et de point de vue. Les résultats d'évaluation de notre méthode à la sous-tâche de localisation de la campagne d'évaluation TRECVID 2013 sont encourageants pour la suite des travaux.
Dans le futur, certaines modifications pourraient être apportées pour améliorer la localisation, par exemple réitérer plusieurs fois la méthode de détection du cadre englobant uniquement sur la zone délimitée précédemment par un cadre englobant afin d'affiner le dessin du cadre englobant. Nous pourrons également modifier la méthode proposée afin de la rendre capable d'extraire plusieurs cadres englobants dans une même image pour gérer les cas où l'objet apparait plusieurs fois dans l'image.

5

Classification des plans de vidéos à partir d'annotations globales niveau vidéos

Dans ce chapitre, nous proposons deux méthodes pour réduire le bruit causé par des annotations non exactes au niveau des plans obtenues par une projection des annotations au niveau de la vidéo sur les plans. Notre première méthode s'emploie à faire le tri dans l'ensemble des plans des vidéos positives pour enlever un maximum de plans négatifs (faux positifs) et garder un maximum de plans positifs. Notre deuxième méthode attribue des poids plus ou moins importants aux plans des vidéos positives et effectue un apprentissage pondéré sur l'ensemble des plans.
Les deux méthodes ont été implémentées et évaluées sur deux collections de données différentes HLF-TRECVid 2008 et MED-TRECVid 2011.

5.1 Motivations : Réduction du bruit causé par des annotations non exactes

La reconnaissance et la détection de concepts ou d'événements dans les documents vidéos, se fait par apprentissage supervisé ou non supervisé. Les informations contenues dans les vidéos sont, tout d'abord, représentées par des descripteurs globaux ou locaux, puis une phase d'apprentissage est réalisée. Dans le cas d'apprentissage non supervisé, les vidéos de l'ensemble d'apprentissage ne sont pas annotées (étiquetées) et le système d'apprentissage tente de trouver des structures cachées entre les données non étiquetées et leurs descripteurs en les regroupant dans des classes non-nommées. Dans le cas d'apprentissage supervisé, les vidéos d'apprentissage sont étiquetées manuellement et le système d'apprentissage analyse l'ensemble des descripteurs des vidéos et des étiquettes attribuées pour générer un modèle capable de prédire l'étiquette d'une nouvelle vidéo. À cause des problèmes de variabilité de forme, de position, de point de vue et d'éclairage dans les vidéos (abordés dans l'introduction) l'apprentissage supervisé reste la méthode la plus efficace pour la reconnaissance et la détection de concepts ou d'événements dans les documents vidéos.

La clé de la réussite des systèmes de détection de concepts dans les vidéos par apprentissage supervisé est la disponibilité de données d'apprentissage annotées adéquatement c'est-à-dire par des annotations au niveau de segments vidéos dont le contenu est homogène. C'est par exemple le cas pour les données de la tâche d'indexation sémantique de TRECVid où les vidéos sont annotées au niveau des plans. L'annotation est une étape manuelle, longue et coûteuse, et la majorité des jeux de données disponibles actuellement sont annotées uniquement au niveau de la vidéo entière malgré l'hétérogénéité de son contenu. Par conséquent, certains plans de la vidéo peuvent ne pas être cohérents visuellement avec l'étiquette attribuée globalement à la vidéo. En théorie, ce dernier problème se résout en calculant des descripteurs globaux au niveau de la vidéo mais en pratique cette solution est mauvaise car l'hétérogénéité de la vidéo dilue le contenu pertinent et le rend difficilement représentable par les descripteurs. Une deuxième solution possible serait de projeter les annotations au niveau de la vidéo sur les plans. La projection de l'étiquette des

vidéos négatives ne provoque pas de problèmes particuliers dans les sacs négatifs (l'ensemble des plans des vidéos négatives). Par contre la projection de l'étiquette des vidéos positives entraîne des problèmes de projection ambiguë en annotant positivement un plan non relié visuellement à l'étiquette et rajoute potentiellement du bruit avec de faux positifs dans les sacs positifs (l'ensemble des plans des vidéos positives).

Pour résoudre le problème des annotations inexactes qui dégradent la performance des systèmes de classification, plusieurs méthodes ont été proposées. Ulges *et al.* ont proposé une méthode probabiliste pour apprendre sur des vidéos d'apprentissage en présence d'images non pertinentes [Ulge 08]. Ils ont modélisé la pertinence de chaque image sous la forme d'une variable aléatoire latente dont la valeur est estimée durant l'apprentissage. Gu *et al.* ont traité le problème comme un cas d'apprentissage d'instance multiples (MIL) où les données d'apprentissage sont regroupées dans des sacs d'échantillons et où chaque sac peut contenir des images non pertinentes en plus de celles qui le sont [Gu 08]. Ils ont groupé les images des vidéos dans des sacs d'échantillons et proposé une fonction de noyau pour apprendre de ces sacs. Habibian *et al.* ont proposé une méthode simple pour détecter les images ne contenant pas d'information visuelle reliée à l'étiquette qui lui est attribuée, dites images vides (ou stop-frames) [Habi 14b]. Ils ont identifié ces images comme étant les images mal classées par plusieurs classificateurs sémantiques.

La motivation initiale de notre travail est de réduire le bruit causé par des annotations non exactes au niveau des plans obtenues par une projection des annotations au niveau de la vidéo sur les plans afin d'améliorer la qualité des systèmes dans le contexte de la tâche MED de TRECVid et qui serait éventuellement applicable à d'autres cas. En effet, cette tâche de TRECVid pose la problématique de la détection des événements dans les vidéos en ne fournissant que des annotations au niveau de la vidéo. Notre première idée est de grouper les plans des vidéos dans des sacs d'échantillons et de faire le tri dans les sacs positifs pour enlever un maximum d'échantillons négatifs (faux positifs) et garder un maximum de vrais positifs dans le but d'obtenir une meilleure pureté des sacs positifs (voir la section 5.2). Notre seconde idée est d'effectuer un apprentissage pondéré en attribuant des poids plus ou moins importants aux échantillons des sacs positifs. Les poids sont attribués selon la méthode détaillée dans la section 5.3.
Nos deux idées sont inspirées de travaux effectués sur le reclassement d'images obtenues par un moteur de recherche suite à une recherche textuelle en fonction de leur caractéristiques visuelles [Quen 12]. Ce reclassement d'images a permis d'améliorer les résultats de recherche en mettant les images les plus pertinentes en tête de classement. Quénot *et al.* se basent sur l'idée que les images pertinentes doivent être semblables entre elles et que les images non pertinentes doivent être différentes entre elles et différentes des images pertinentes. Ils ont implémentés l'idée en classant les images en fonction de la distance moyenne de celles-ci avec leurs plus proches voisines. Les méthodes que nous proposons ici classent les plans des vidéos positives

par ordre croissant d'un certain score de distance (ou densité) par rapport à leurs plus proches voisins, ensuite elles retirent des sacs positifs les plans dont le score dépasse un certain seuil s (tout en prenant le risque de perdre quelques vrais positifs) et/ou pondèrent les plans selon ces scores.

Le reste de ce chapitre sera organisé comme suit : dans la section 5.2 nous détaillerons notre première idée basée sur le tri des annotations ainsi que son évaluation sur deux collections de données différentes ; dans la section 5.3 nous aborderons une idée alternative basée sur l'apprentissage automatique et son évaluation sur les mêmes deux collections de données. Nous proposons dans les deux sections une analyse approfondie des résultats obtenus.

5.2 Production automatique de nouvelles annotations

Pour la mise en place de notre première idée basée sur le tri des sacs positifs pour enlever un maximum d'échantillons négatifs (faux positifs) et garder un maximum d'échantillons de vrais positifs, nous proposons une approche qui permet de produire automatiquement une nouvelle annotation au niveau des plans à partir d'une annotation au niveau de la vidéo. Cette annotation au niveau des plans est effectuée en se basant sur le contenu des images des plans et un contenu qui peut être visuel, textuel ou audio.

En d'autres termes, nous cherchons à générer des annotations plus précises au niveau des plans à partir des annotations au niveau vidéos (comme illustré dans la figure 5.1) pour améliorer la qualité des modèle d'apprentissage et la performance des système de classification.

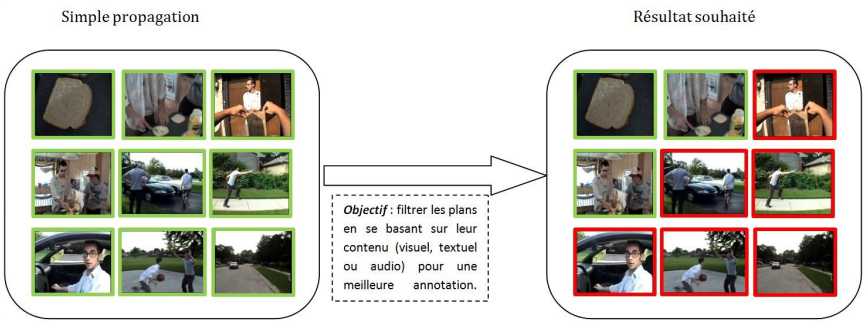


FIGURE 5.1 – Une illustration de l'objectif final de notre méthode.

5.2.1 Modèle proposé

Nous détaillons ici la méthode que nous proposons pour générer automatiquement une annotation au niveau des plans à partir d'une annotation au niveau de la vidéo. La méthode peut être considérée comme un post-traitement des données d'entraînement

car elle intervient en amont du processus d'apprentissage d'un système d'indexation. Elle est inspirée des travaux effectués par Quénot *et al.* qui ont permis d'améliorer le reclassement d'images obtenues par un moteur de recherche [Quen 12]. Les plans étant représentés dans un espace de description, la méthode est fondée sur les hypothèses non-indépendantes suivantes :

($h1$) La similarité entre les plans représentant un événement donné (plans positifs) doit être supérieure à celle entre les plans représentant un événement donné (plans positifs) et les plans ne représentant pas ce même événement (plans négatifs).

($h2$) Les représentations des plans positifs doivent être regroupées alors que celles des plans restants doivent être éloignées et dispersées.

($h3$) Un plan positif doit avoir statistiquement une distance moyenne à ses voisins les plus proches plus importante qu'un plan négatif ou doit se trouver dans une région de plus fore densité.

Nous nous sommes concentrés uniquement sur les vidéos positives (sacs de plans positifs), notre méthode ne traite donc que les plans des vidéos positives selon les hypothèses citées ci-dessus. C'est une méthode en trois étapes détaillées ci-dessous et illustrées dans la figure 5.2.

5.2.1.1 Matrice de distance

Soit un descripteur x et une distance d entre descripteurs, nous calculons une matrice de distance M de taille $N \times N$ où N est le nombre de plans initialement dans les sacs positifs. La matrice M est la matrice de distance entre chaque paire de plans des vidéos positives. Cette distance d sera calculée sous la forme d'une distance euclidienne.

5.2.1.2 Distance représentative

Une fois que la matrice de distance est calculée, nous calculerons une distance globale D intégrant les distances d'un plan à un ensemble de voisins. Chaque plan sera représenté par cette distance globale. Il existe différentes façons de calculer cette distance globale :

- La première est la distance moyenne d'un plan x_i à ses k plus proches voisins :

$$D(x_i) = \frac{\sum_{j=1}^{j=k} d(x_i, x_{n(i,j)})}{k}$$

où $n(i,j)$ est l'indice du $j^{\text{ième}}$ voisin le plus proche du plan d'indice i.

- La seconde est la distance médiane entre un plan x_i et ses k plus proches voisins qui est aussi la distance au $(\frac{k}{2})^{i\text{ème}}$ voisin.

- Enfin, on peut considérer une distance pondérée entre un plan x_i et ses k plus proches voisins, de façon à prendre en compte le degré d'éloignement des voisins :

$$D(x_i) = \frac{\sum_{j=1}^{j=k} f(j) d(x_i, x_{n(i,j)})}{\sum_{j=1}^{j=k} f(j)}$$

  k et f sont des paramètres définissant une variante donnée de la fonction D. Le choix de la distance entre descripteurs est à déterminer par validation croisée sur l'ensemble de développement.

5.2.1.3 Choix du seuil

La dernière étape consiste à fixer un seuil à partir des distances globales calculées et c'est à partir de ce seuil que nous générerons les nouvelles annotations des plans des vidéos positives. Si la distance globale d'un de ces plans dépasse le seuil fixé, le plan sera considéré loin ou différents des autres plans par la suite il sera ignoré. Là encore le seuil peut être calculé de différentes façons :

- Moyenne : le seuil sera calculé à partir d'une moyenne sur les distances globales.

- Médiane : le seuil sera calculé à partir d'une médiane sur les distances globales.

- Pourcentage : le seuil sera un pourcentage p de façon à ne garder que $p\%$ des plans positifs les plus proches les uns des autres.

5.2.2 Evaluation sur MED-TRECVid 2011

Nous avons voulu évalué la justesse de ces hypothèses et la mesure dans laquelle les nouvelles annotations peuvent effectivement améliorer l'apprentissage. Pour ce faire, nous avons utilisé la collection de vidéos produite par TRECVID pour la tâche de détection des événements (MED) 2011.

5.2.2.1 La collection de données

Cette collection contient des 40 000 vidéos collectées d'Internet et annotées pour 15 événements complexes au niveau vidéo. Dans les événements annotés on trouve : *Attempting a board trick, Feeding an animal, Landing a fish, Wedding ceremony, Working on a woodworking project, Birthday party, Changing a vehicle tire, Flash mob gathering, Getting a vehicle unstuck, Grooming an animal, Making a sandwich, Parade, Parkour, Repairing an appliance, Working on a sewing project.*
La collection est divisée en deux parties, une partie développement contenant 13 115

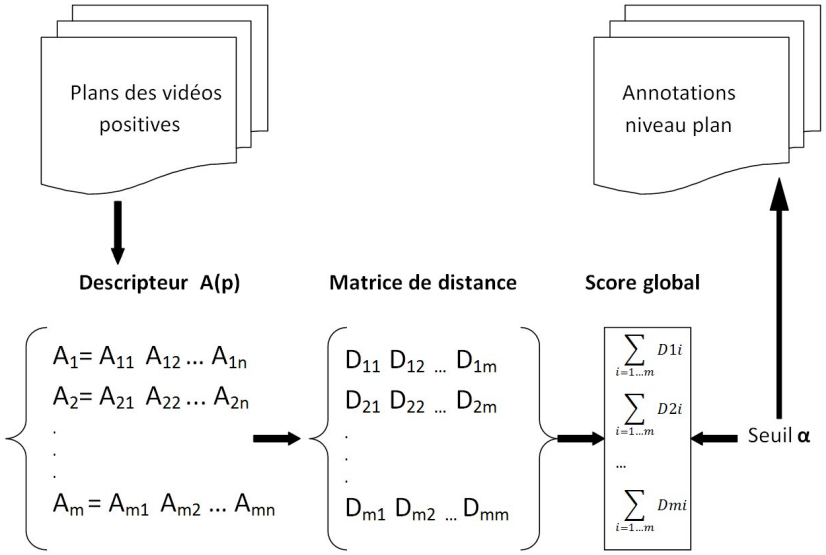


FIGURE 5.2 – Les trois étapes de l'approche proposée.

vidéos pour un total d'environ 370 heures et une partie test contenant 32 061 vidéos dontla durée totale est d'environ 1200 heures.

Pour l'évaluation de notre approche, nous comparons la qualité de la détection automatique d'événements entre un système entraîné avec des fichiers d'annotations niveau plan généré en propageant les annotations des vidéos positives à tous leurs plans respectifs avec celui entrainé avec les fichiers d'annotations produits par notre approche. La mesure appropriée pour l'estimation de la qualité de détection est la précision moyenne (AP). Dans notre cas, nous disposons de plusieurs evénements donc la mesure de qualité globale sera la MAP.

5.2.2.2 Résultats obtenus

Nous avons comparé les différentes variantes de notre approche entre elles et avec la méthode classique de référence qu'on appellera la « baseline » (annotations propagées simplement du niveau vidéo au niveau plan). Les expérimentations utilisent un descripteur combinant la couleur et la texture construit par des « mots visuels » sur le descripteur de couleur et de texture optimisé le « hg104pw0.300p52 » [Dele 11] . Nous avons relancé un apprentissage avec les nouvelles annotations générées par notre modèle et comparé les capacités de détection des événements de ce modèle avec celle de la baseline. Dans un premier temps, nous avons testé sur les données de développement les différentes variantes de l'approche. Ces variantes considèrent les différentes méthodes possibles pour calculer les distances représentatives ainsi que pour le choix du seuil. Les meilleurs résultats sont obtenus par une distance pondérée comme distance représentative et un seuil en pourcentage.

La figure 5.3 reprend les résultats obtenus par notre méthode en faisant varier le

nombre K plus proche voisins pour le calcul de la distance représentative et en faisant varier en même temps le pourcentage pour le seuil de sélection des plans. Comme on peut voir dans cette figure la meilleure performance est obtenue avec $K = 70$ et avec un seuil à 80%. La MAP atteint 0.1743, soit une amélioration relative de 1.3%.

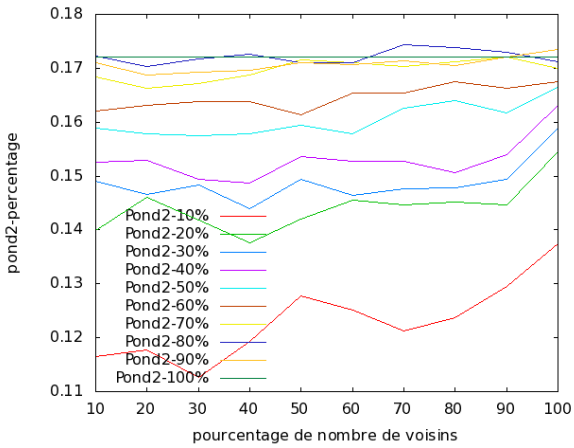


FIGURE 5.3 – Les performances de l'approche en variant le paramètre K et le pourcentage

Contrairement à nos attentes, cette amélioration reste faible. Nous avons donc analysé les résultats concept par concept. Nous avons pu constater que l'amélioration pour certains concepts comme « Birthday party » peut être beaucoup plus élevée que pour d'autres comme « Parkour » comme on peut l'observer dans les figures 5.4 et 5.5. Une explication possible est que pour des concepts comme « Parkour » ou encore « Parade » notre approche ne sera pas bien utile car ces événements durent tout le long de la vidéo. Pour d'autres comme « Making a sandwich » l'événement est plutôt de courte durée et ne se produit que durant quelques minutes dans une vidéo qui est bien plus longue et variée.

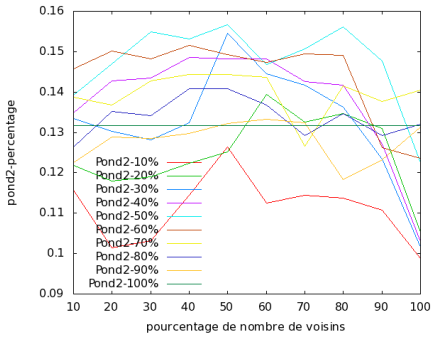


FIGURE 5.4 – Les performances pour le concept « Birthday Party ».

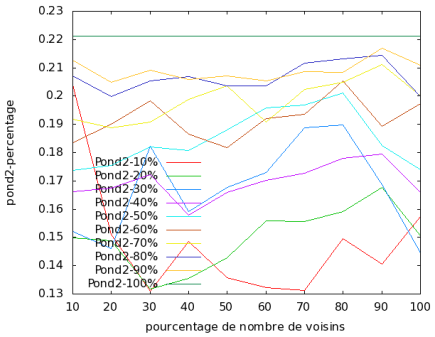


FIGURE 5.5 – Les performances pour le concept « Parkour ».

Dans une expérience complémentaire, nous avons testé le filtrage sur les données de test de la collection de MED 2011 avec des paramètres optimisés concept par

	Développement		Test	
Evénements	Baseline	Filtrage	Baseline	Filtrage
Birthday Party	0.1316	0.1567	0.0084	0.0085
Changing a vehicle tire	0.0438	0.0570	0.0092	0.0067
Flash mob gathering	0.5248	0.5658	0.0326	0.0305
Getting a vehicle unstuck	0.1910	0.1910	0.0296	0.0296
Grooming an animal	0.0965	0.1112	0.0122	0.0108
Making a sandwich	0.1019	0.1176	0.0054	0.0062
Parade	0.1773	0.1912	0.0639	0.0569
Parkour	0.2212	0.2212	0.0034	0.0034
Repairing an appliance	0.4099	0.4665	0.0149	0.0070
Working on a sewing project	0.1079	0.1533	0.0277	0.0283

TABLE 5.1 – L'AP par événements sur les données du développement et test

concept sur les données du développement. Le tableau 5.1 reprend les résultats obtenus par événement sur les données de développement et sur celles de test. Nous constatons de bonnes ou très bonnes améliorations sur les données de développement pour huit concepts sur les dix évalués mais seulement de légères améliorations dans les données de test sur seulement trois concepts. Ceci nous amène à conclure que la méthode ne se généralise pas bien au niveau de tous les concepts.

Le descripteur couleur et texture a été choisi pour sa simplicité et aussi parce qu'il a été montré qu'il constituait une bonne « baseline » pour l'indexation sémantique des vidéos [Dele 11]. D'un autre côté, ce descripteur a pu induire de l'instabilité dans notre approche car il se base seulement sur la couleur et sur la texture pour représenter le contenu. Ces critères ne sont pas forcément les plus adaptés pour détecter des événements. D'autres descripteurs comme les SIFT [Lowe 04] ou des descripteurs basés sur le mouvement comme les STIP [Lapt 05] donneront peut-être de meilleurs résultats car ils représenteront un contenu plus spécifique aux événements que la couleur et la texture. Enfin, bien que cette méthode soit applicable à tous types de concepts et d'événements, sa performance mitigée sur les dix différents concepts est éventuellement due à la nature des concepts à détecter dans le cadre de MED qui sont particulièrement complexes.

5.2.3 Évaluation sur HLF-TRECVid 2008

Pour mieux comprendre les résultats précédents, nous avons voulu évaluer la précision des annotations au niveau des plans générées par notre méthode : pour ce faire, nous avions besoin d'une collection de données entièrement annotées au niveau de plans et nous avons choisi la collection de vidéos de la tâche High Level Feature (HLF) de TRECVID 2008.

5.2.3.1 La collection de données

Cette collection contient 219 vidéos découpées en 43 616 plans et annotées pour 30 concepts. Parmi les concepts annotées, certains sont dynamiques et d'autres sont statiques : *Airplane flying, Boat ship, Bridge, Bus, Chair, Cityscape, Classroom, Demonstration or protest, Dog, Doorway, Drive, Emergency vehicle, Female human face closeup, Flower, Hand, Harbor, Infant, Kitchen, Mountain, Nighttime, People dancing, Person eating, Person playing a musical instrument, Person riding a bicycle, Singing, Street, Telephone, Traffic intersection, Two people*.

La collection est divisée en deux parties, une partie développement contenant 110 vidéos pour 21 532 plans et une partie test contenant 109 vidéos pour 22 084 plans.

5.2.3.2 Résultats obtenus

Les expérimentations utilisent le descripteur combinant la couleur et la texture construit par des « mots visuels » sur le descripteur de couleur et de texture optimisé le « hg104 » [Dele 11] . Nous avons utilisé la distance euclidienne pour la calcul de la matrice de distance entre tous les plans des vidéos positives. Pour le calcul du score global de chaque plan nous avons testé différentes valeurs de k, $k = 10\%, 20\%, 50\%, 100\%$. Enfin pour le choix du seuil, nous avons opté pour un seuil en pourcentage ($p\%$) que nous avons fait varier entre 10% et 100% ($p = 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%$).

La figure 5.6 montre la courbe du rappel×précision obtenue en triant les plans avec notre méthode avec différentes valeurs du seuil p avec la méthode classique de projection des annotations du niveau vidéo au niveau des plans (dont le résultats est obtenu avec notre méthode mais en sélectionnant tous les plans des vidéos positives donc avec le seuil $p = 100\%$). La courbe représente alors la courbe du rappel×précision pour les 30 concepts et avec les 10 différentes valeurs de $p = 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, 90\%, 100\%$. Nous constatons que plus le seuil en pourcentage (p) augmente plus la performance baisse, donc plus le nombre de plans considérés comme positif augmente plus la qualité du système d'indexation devient mauvaise. Ceci montre l'importance de la sélection des régions de densité pour l'annotation des plans.

Le tableau 5.2 montre la précision@N des annotations au niveau de plans produites par notre méthode. Nous remarquons que notre méthode a réussi à classer plus de plans réellement positifs en haut de la liste des plans produite.

5.2.4 Analyse des résultats

Nous avons présenté une méthode permettant de produire des annotations au niveau des plans afin de réduire le bruit causé par une simple projection des annotations des vidéos au niveau des plans dans le contexte d'un apprentissage supervisé pour la détection de concepts ou d'événements. La méthode proposée utilise le contenu visuel des plans et des vidéos en se basant sur l'idée que les plans contenant

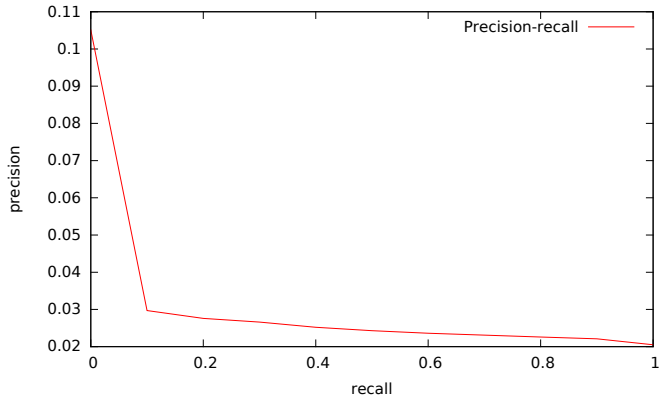


FIGURE 5.6 – la courbe du rappel×précision obtenue en triant les plans avec la méthode proposée sur les données de la tâche HLF de TRECVid 2008.

un concept ou un événement sont semblables alors que les plans ne représentant pas ce concept ou cet événement sont différentes entre eux et des plans qui les contiennent. L'approche a été implémentée et évaluée dans le cadre de la tâche de détection d'événements multimédia (MED) de TRECVID 2011 en produisant automatiquement des annotations au niveau des plans et en évaluant ensuite la capacité de détection des événements des modèles ayant été entraînés sur ces données. Cette approche n'a pas toujours permis d'améliorer les performances de la détection d'événements mais elle a tout de même apporté des améliorations pour quelques événements. Nous avons évalué également la qualité des annotations générées sur la collection de données de la tâche HLF de TRECVid 2008. Les résultats ont montrés que notre méthode est capable de classer plus de plans vrais positifs en haut de la liste des plans.

Malgré la capacité de notre méthode à retrouver des vrais plans positifs (comme montré dans la section 5.2.3), la performance du système d'indexation a été mitigée (comme montré dans la section 5.2.2), peut-être que notre méthode enlève des plans vrais positifs en essayant d'enlever les plans faux positifs (négatifs) ce qui nuit à la qualité du modèle d'apprentissage et entraîne ses performances.

5.3 Pondération des plans des vidéos d'entraînement

Devant l'échec de la première idée qui conduit à enlever des échantillons vrais positifs en essayant d'enlever le maximum d'échantillons faux positifs (i.e. des plans négatifs) contenus dans les sacs positifs, nous avons décidé de pondérer ces échantillons plutôt que de les enlever complètement en les retirant ou en les annotant comme négatifs. Le poids attribué à chaque plan des vidéos d'apprentissage positives correspondra à la probabilité estimé pour un plan d'être un vrai positif.

Pour le calcul de la probabilité d'un plan d'être un vrai positif, donc du poids

P@N	P
P1	0.0417
P2	0.0333
P5	0.0311
P10	0.0313
P15	0.0247
p20	0.0225
P100	0.0213
P200	0.0224
P500	0.0226
P1000	0.0231
P2000	0.0225
P5000	0.0215
P10000	0.0166
P20000	0.0108
Pall	0.0099

TABLE 5.2 – Precision à N pour la méthode proposée sur les données de la tâche HLF de TRECVid 2008.

qui lui sera attribué, nous nous sommes basés sur les mêmes hypothèses ($h1, h2, h3$) que la méthode précédente (section 5.2.1). Concrètement (h_1) est traduite par : plus un plan est proche des autres plans des vidéos positives et loin des plans des vidéos négatives, plus le poids du plan sera élevé. Mathématiquement, nous avons opté pour un calcul de densité qui prend en considération les distances entre les différents plans pour décider du poids à attribuer à un plan.

Nous soulignons que dans la méthode précédente nous n'avions pris en compte que la moitié des hypothèses en ne considérant que les vidéos positives pour le tri des sacs positives (section 5.2.1), pour l'apprentissage pondéré nous avons voulu prendre en compte l'intégralité des hypothèses. Par conséquent, nous proposons deux modèles pour la pondération des plans : le premier est basé sur un calcul de densité entre les échantillons positifs de l'ensemble d'entraînement et le second est basé sur un calcul de densité entre les échantillons positifs et les échantillons négatifs de l'ensemble d'entraînement. Ces deux modèles sont détaillés ci-dessous.

5.3.1 Pondération à partir des vidéos positives

Dans ce premier modèle, les poids attribués aux plans des vidéos d'entraînement positives, nécessaires pour la phase d'apprentissage pondéré, sont calculés à partir des vidéos positives. Le modèle comporte deux étapes :

Matrice de distance Soit un descripteur x et une distance d entre descripteurs, nous calculons une matrice de distance M de taille $N \times N$ où N est le nombre de plans des vidéos annotées positives. La matrice M est la matrice de distance entre

chaque deux plans des vidéos positives. Cette distance d sera calculée sous la forme d'une distance euclidienne comme précédemment.

Poids Une fois la matrice de distance calculée, nous calculons le score de densité (ou le poids) D_p pour chacun des plans et qui sera pris en compte lors de l'apprentissage. D_p intègre la distance d'un plan par rapport à l'ensemble de ses voisins et elle est calculée par la formule suivante pour un plan x_i et ses k plus proches voisins :

$$D_p(x_i) = \frac{\sum_{j=1}^{j=k} \exp(-\frac{d'^2}{2})}{k} \tag{5.1}$$

et

$$d' = \frac{d(x_i, x_{n(i,j)})}{\sigma \times d_m} \tag{5.2}$$

où $n(i,j)$ est l'indice du $j^{\text{ème}}$ voisin le plus proche du plan x_i et d_m correspond à la distance moyenne statistique.

5.3.2 Pondération à partir des vidéos positives et négatives

Dans ce second modèle, les poids attribués aux plans des vidéos d'entraînement positives, et nécessaires pour la phase d'apprentissage pondéré, sont calculés à partir des vidéos positives et des vidéos négatives. Pour cela, nous avons utilisé un modèle basé sur le principe des k plus proches voisins (kNN). Le score de densité (ou le poids) D_{pn} attribué pour chacun des plans est obtenu selon les formules suivantes :

$$D_{pn}(x_i) = \frac{C_p}{C_n + C_p} \tag{5.3}$$

et

$$C_p = (\frac{1}{freq_p})^\alpha \times \sum_{lab_{x_i}=p} (\exp -\frac{(x - x_i)^2}{2\sigma^2 d_m^2}) \tag{5.4}$$

$$C_n = (\frac{1}{freq_n})^\alpha \times \sum_{lab_{x_i}=n} (\exp -\frac{(x - x_i)^2}{2\sigma^2 d_m^2}) \tag{5.5}$$

où d_m correspond à la distance moyenne statistique et $(\frac{1}{freq_n})^\alpha$ est un terme pour équilibrer les classes comme c'est utilisé en classification kNN.

5.3.3 Résultats obtenus sur HLF-TRECVid 2008

Nous avons évalué notre méthode sur deux collections de données : TRECVid High Level Frequency (HLF) 2008 et TRECVid Multimedia Event Detection (MED) 2011. Nos premières expérimentations sur MED 2011 n'ont une fois encore pas montré de gain ou aucun statistiquement significatif. Par soucis de clarté, nous ne détaillons ici que les expérimentations effectuées sur TRECVid HLF 2008 (une collection de

données présentée précédemment dans 5.2.3.1).

Pour évaluer la qualité de nos modèles de pondération, nous avons comparé la performance de quatre systèmes d'indexation entraînés de différentes manières :

- Le premier entraîné avec les annotations au niveau des plans fournies par TRECVid (S_1),
- Le second entraîné avec les annotations projetées du niveau des vidéos seulement (ou sacs de plans), le travail décrit dans le chapitre (S_2),
- Le troisième entraîné avec les annotations projetées du niveau de la vidéo au niveau du plan et pondérées selon D_p (cf. 5.3.1) (S_3),
- Le quatrième entraîné avec les annotations projetées du niveau de la vidéo au niveau du plan et pondérées selon D_{pn} (cf. 5.3.2) (S_4).

Nous précisions que la métrique d'évaluation utilisée dans l'évaluation des système d'indexation est la moyenne de la précision moyenne (MAP). Lors de nos expérimentations sur l'apprentissage pondéré, nous avons attribué aux plans des vidéos positives des poids calculés par les méthodes proposées alors que pour les plans des vidéos négatives nous avons fixé leur poids à 1.0. De plus nous avons testé différentes valeurs des variables : K entre 10% et 100%, σ entre 0.09 et 3.0 et enfin nous avons appliqué une normalisation sur l'ensemble des poids avec un coefficient n où $n = 0.5, 1, 1.5$.

Globalement, nous avons constaté que la performance du système d'indexation S_3 augmentait en même temps que σ augmentait jusqu'à atteindre un maximum à 0.0147 alors que k et n ne l'impactaient pas. Et les résultats obtenus par le système S_3, appris sur les données pondérées par notre modèle avec les différentes variantes, ont très peu de variabilité avec une MAP toujours comprise entre 0.0138 et 0.0147. De plus, la performance du système S_4, qui prend en considération les plans positifs et négatifs pour la pondération, ne dépasse pas celle de S_3.

Le tableau 5.3 rapporte la performance des trois systèmes d'indexation : S_1, S_2, S_3 où les poids sont calculés par notre méthode (décrite dans la section 5.3.1) avec $k = 10\%$, $n = 0.5$ et $\sigma = 0.90$. Comme attendu, S_1 obtient les meilleurs résultats et dépasse largement la performance de S_2 et S_3 avec une $MAP = 0.0607$. Cette performance est logique sachant que les annotations des données d'apprentissage de S_1 sont celles fournies par TRECVid et qu'elles sont effectuées manuellement au niveau des plans, par conséquent elles contiennent très peu de bruit. En ce qui concerne S_3, globalement S_2 et S_3 ont donné les mêmes résultats avec une MAP de 0.0147, ce qui montre que notre méthode de pondération n'a pas réussi à diminuer le bruit dans les données d'apprentissage avec des annotations projetées du niveau vidéos au niveau plan malgré les améliorations qu'elle a réussi à apporter pour certains concepts

Concepts	Vérité terrain (S_1)	Propagation (S_2)	Pondération (S_3)
Airplane_flying	**0.0567**	0.0070	0.0060
Boat_Ship	**0.1060**	0.0160	0.0140
Bridge	**0.0132**	0.0031	0.0134
Bus	**0.0071**	0.0011	0.0012
Cityscape	**0.0465**	0.0090	0.0089
Classroom	**0.0177**	0.0058	0.0047
Demonstration_Or_Protest	**0.0169**	0.0161	0.0170
Dog	**0.0957**	0.0034	0.0027
Driver	**0.0840**	0.0104	0.0093
Emergency_Vehicle	**0.0044**	0.0008	0.0008
Flower	**0.0727**	0.0080	0.0120
Hand	**0.1064**	0.0355	0.0367
Harbor	**0.0733**	0.0072	0.0072
Kitchen	**0.0096**	0.0032	0.0028
Mountain	**0.0273**	0.0073	0.0065
Nighttime	**0.1628**	0.0107	0.0105
Singing	**0.0264**	0.0132	0.0143
Street	**0.1337**	0.0304	0.0296
Telephone	**0.0224**	0.0076	0.0070
Two_people	**0.1303**	0.0986	0.0985
all	**0.0607**	0.0147	0.0147

TABLE 5.3 – Comparaison de la précision moyenne obtenue par les des trois systèmes d'indexation (S_1, S_2, S_3) pour chacun des concepts.

(comme Flower, Hand, Singing).

5.3.4 Analyse des résultats

Pour mieux comprendre les raisons pour lesquelles la pondération des plans n'a pas permis l'amélioration de performance espérée par l'intermédiaire de la réduction du bruit, nous avons analysé la distribution des scores attribués par nos modèles et la répartition des vidéos du corpus utilisé. Idéalement, la courbe représentant les scores attribués aux plans vrais positifs et celle représentant les scores attribués aux faux positifs doivent être complètement séparées ou avec très peu d'intersection. Or il s'est avéré que ce n'est pas le cas en pratique. Les figures 5.7 et 5.8 montrent deux exemples représentatifs des répartitions des poids normalisés attribués par notre modèle (décrit dans la section 5.3.1) aux plans des vidéos positives et analysées séparément pour les plans vrais positifs et les plans négatifs (faux positifs). Comme nous pouvons le remarquer les deux courbes se chevauchent en grande partie ce qui montre que notre modèle ne réussit pas à différencier les faux positifs des vrais positifs et il leur attribue des scores similaires. Par conséquent, si on pondère ou on filtre

(5.2) selon ces scores au niveau de la première intersection entre les deux courbes un grand nombre de plans faux positifs seront éliminés mais également un nombre considérable de plans vrais positifs.

Nous avons également analysé les données de la collection TRECVd HLF 2008. Nos analyses ont montré que le nombre des vidéos positives pour un concept donné est faible ($\sim 20\%$) et que la proportion des plans vrais positifs dans l'ensemble des plans des vidéos positives (sacs positifs) est encore plus faible et atteint en moyenne 2% (soit seulement 2% des plans des vidéos positives contiennent le concept). Les plans vrais positifs se retrouvent donc noyés dans l'ensemble des plans négatifs qui sont bien plus fréquents, ce qui fausse le calcul des densités et l'attribution des poids. Pour réduire le problème de proportions, nous avons re-découpé les vidéos de la collection TRECVd HLF 2008 en vidéos plus courtes de manière à augmenter le nombre de plans vrais positifs, et avoir 10 plans par vidéos et donc passer de 2% à 20% de plans vrais positifs.
Comme les figures 5.9 et 5.10 le montrent, la répartition des poids calculés selon le nouveau découpage des vidéos (5.10) est légèrement meilleure que celle des poids calculés selon le découpage initial des vidéos où la proportion des plans vrais positifs est très faible (5.10). En effet, avec le nouveau découpage augmentant la proportion des plans vrais positifs dans les données d'apprentissage, notre modèle a réussi à attribuer plus de poids élevés aux plans vrais positifs qu'aux plans faux positifs. Ainsi un faible gain a été constaté sur la performance du système d'indexation par pondération selon le nouveau découpage des vidéos d'apprentissage.

Enfin, la méthode proposée utilise le contenu visuel des plans et un calcul de la densité entre plans, en se basant sur l'idée que les plans contenant un concept ou un événement sont proches et forment des regroupements alors que les plans ne représentant pas ce concept ou cet événement sont différents entre eux et du reste des plans et donc loin des regroupements. L'approche a été implémentée avec différentes variantes et évaluée dans le cadre de la tâche de détection d'événements multimédia (MED) de TRECVID 2011 et de la tâche HLF de TRECVid 2008. La rareté des plans vrais positifs (2% des vidéos positives) dans les données d'apprentissage a compliqué le calcul de densité et faussé l'attribution des poids aux différents plans. Par conséquent, la pondération des échantillons n'a pas permis une améliorations de la performance des systèmes d'indexation. Néanmoins, le re-découpage des plans a montré un faible gain. Á noter en comparaison, dans le travail [Quen 12] qui a inspiré cette étude, la proportion de vrais positifs dans les sacs positifs était bien plus favorable (50-60%).

5.4 Conclusion

Dans ce chapitre nous avons présenté deux différentes idées pour réduire le bruit produit par des annotations non-exactes au niveau des plans, dans le contexte d'un apprentissage supervisé pour la détection de concepts ou d'événements. La première

méthode génère de nouvelles annotations au niveau des plans alors que la deuxième attribue des poids aux différents plans. Nous avons évalué ces deux méthodes dans le cadre de la tâche de détection d'événements multimédia (MED) de TRECVID 2011 et de la tâche HLF de TRECVid 2008.

Les deux idées semblaient intéressantes à priori, mais elles n'ont pas été concluantes malgré de nombreux essais et de différentes variantes dont certaines ont été détaillées précédemment : plusieurs formules pour le calcul de la distance moyenne ou de la densité, utilisation des échantillons négatifs, ...
L'échec des deux méthodes est dû à la rareté des échantillons vrais positifs dans les données d'apprentissage, et donc malgré la tentative d'éliminer les échantillons faux positifs des sacs positives (avec la première méthode 5.2) il subsiste toujours beaucoup d'échantillons négatifs. D'autre part la suppression d'échantillons négatifs a entraîné la suppression d'échantillons vrais positifs alors qu'ils sont au départ très peu fréquents. La rareté de ces plans positifs a faussé le calcul de la densité entre les échantillons et donc l'attribution des poids lors de la deuxième méthode (5.3). Les deux méthodes proposées ne seraient donc pas adaptées aux corpus de vidéos contenant un grand déséquilibre entre les classes.

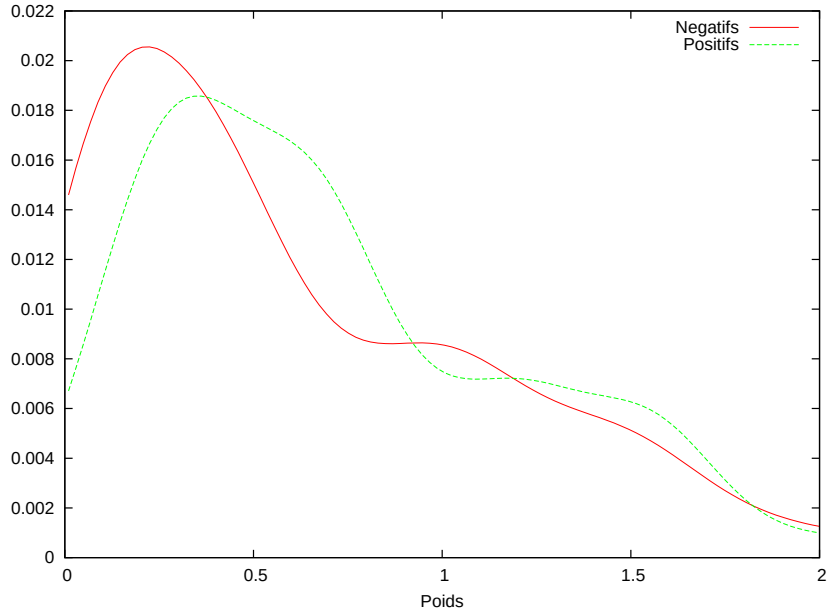


FIGURE 5.7 – La distribution des poids normalisés attribués aux plans des vidéos positives selon le modèle détaillé dans 5.3.1 avec $\sigma = 0,1$.

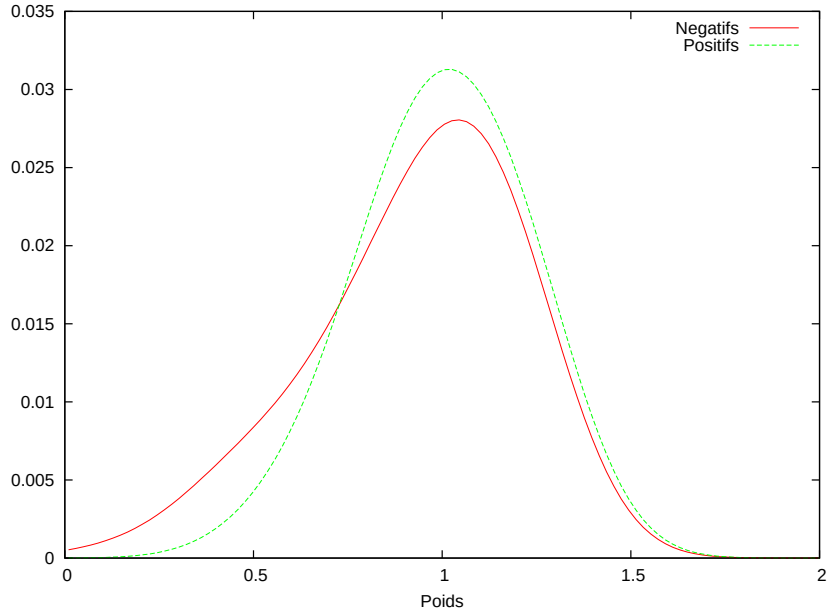


FIGURE 5.8 – La distribution des poids normalisés attribués aux plans des vidéos positives selon le modèle détaillé dans 5.3.1 avec $\sigma = 0,3$.

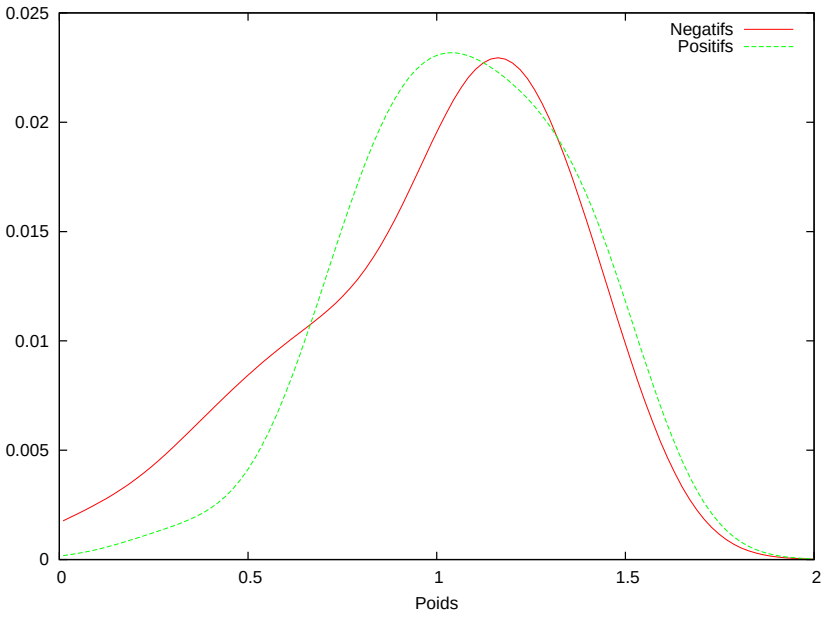


FIGURE 5.9 – La répartition des poids calculés selon le découpage initial des vidéos.

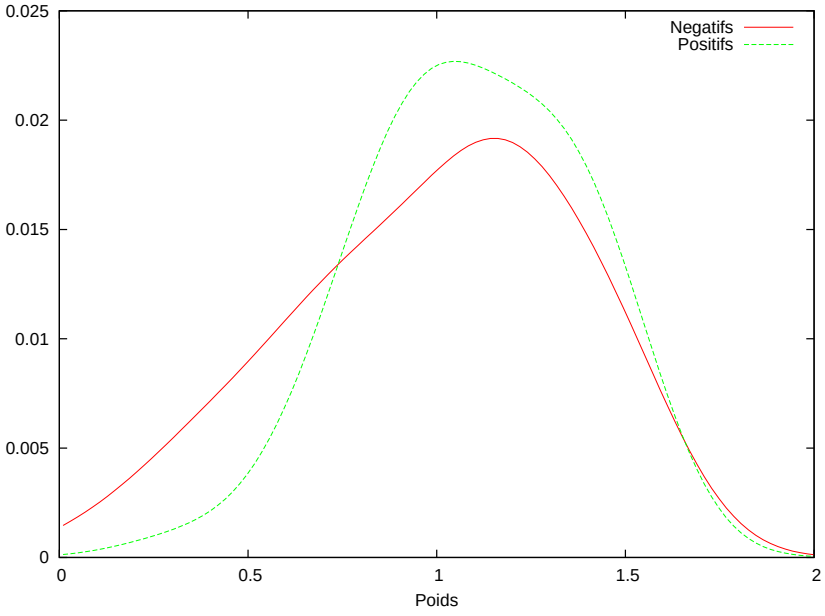


FIGURE 5.10 – La répartition des poids calculés selon le nouveau découpage des vidéos.

6

Optimisation de descripteurs pour l'indexation des documents multimédias

Dans ce chapitre, nous proposons une méthode d'optimisation de descripteurs dédiée aux systèmes d'indexation et de recherche de documents multimédias par le contenu. Une très grande variété de descripteurs existe. Cependant, les descripteurs les plus efficaces possèdent souvent des caractéristiques qui les rendent difficilement utilisables sur les grandes collections de données. Ils sont souvent de très grandes tailles (pouvant aller jusqu'à des centaines de milliers de composantes) et/ou ils ne sont adaptés qu'à des distances très coûteuses en termes de calculs (par exemple la distance χ^2). La méthode proposée combine une transformation non linéaire avant et après une réduction de dimension basée sur une ACP. La transformation obtenue est globalement optimisée. Le descripteur produit est de dimension bien plus réduite que celle du descripteur original tandis que sa performance avec une simple distance euclidienne surpasse celle du descripteur original dans la plupart des cas.
La méthode a été validée et évaluée sur une multitude de descripteurs et en utilisant les jeux de données de la tâche d'indexation sémantique (SIN) de TRECVid 2010. Elle a été appliquée sur l'importante collection de données de la tâche SIN de TRECVId 2012 et sur des centaines de descripteurs de différents types et dont la taille originale varie entre 15 et 32 768 éléments.

6.1 Motivations : Compromis entre performance et dimension des descripteurs

Ces dernières années, beaucoup de travaux de recherche ont visé le développement d'un système d'indexation et de recherche efficace et robuste. Cependant, il reste encore des défis majeurs à relever pour augmenter la performance de ces systèmes, en particulier dans le cas d'applications à grande échelle. La grande majorité de l'état de l'art se focalise sur l'extraction et l'utilisation de descripteurs et néglige l'étape de l'optimisation des descripteurs (2.5). Or les méthodes d'optimisation utilisées jouent un rôle très important dans l'amélioration de la performance de ces systèmes. Elles permettent d'obtenir des vecteurs de descriptions plus compacts et plus précis pour la représentation du contenu. Par conséquent, elles sont capables de réduire significativement le taux d'erreur des systèmes de classification et donc d'augmenter la précision de l'indexation. L'optimisation des descripteurs est une étape cruciale pour les systèmes d'indexation des documents multimédias.

Deux grandes catégories de descripteurs se sont distinguées : les descripteurs globaux et les descripteurs locaux. Les descripteurs locaux capturent plus d'information que les descripteurs globaux, ce qui les rend plus efficaces que ces derniers. Mais une étape d'agrégation des descripteurs locaux est nécessaire pour qu'ils puissent être utilisables par les systèmes de classification. Les méthodes d'agrégation les plus populaires sont : les sacs-de-mots [Csur 04, Sivi 03] (un histogramme de descripteurs locaux est calculé pour constituer un descripteur du contenu global d'une image ou d'un segment vidéo) ou les vecteurs de Fisher [Perr 10b]. Pour plus de détails à propos des descripteurs voir la section 2.4).

Les expériences ont montré que plus la dimension des descripteurs (qu'ils soient sous la forme de sacs-de-mots ou de vecteurs de Fisher) est grande plus ils sont efficaces, cette dimension pouvant atteindre parfois les dizaines de milliers d'éléments. En revanche, cette grande taille implique des problèmes pratiques tels que le temps de calcul ou l'espace de stockage nécessaires. À ces problèmes s'ajoutent ceux reliés à l'apprentissage. En effet, les méthodes populaires d'apprentissage supervisées comme les K-plus proches voisins (KNN) ou les SVM reposent toutes sur un calcul de distance entre les descripteurs, direct ou par l'intermédiaire d'un noyau (comme le noyau RBF). La distance euclidienne et la distance de Chi-deux (χ^2) s'y prêtent très bien mais, plus la dimension des descripteurs est grande, plus le calcul de distance sera long. La distance χ^2 est celle qui est la plus adaptée à la comparaison des descripteurs sous la forme d'histogrammes (comme ceux obtenus par la méthode très populaire des sacs-de-mots) par contre elle possède deux inconvénients : elle est significativement plus coûteuse à calculer et elle n'est pas compatible avec une technique de réduction de dimension basée sur une ACP.

Pour résoudre ces problèmes, Sanchez *et al.* proposent de combiner une méthode d'implémentation efficace (par le produit de quantification) avec une version linéaire du classificateur SVM basée sur une descente stochastique de gradient [Sanc 13]. Une solution alternative serait de réduire la dimension des descripteurs par l'intermédiaire de méthode de réduction de dimension tout en gardant l'efficience d'un classificateur SVM à noyau RBF ou d'un KNN. Ces deux méthodes sont deux solutions très différentes pour résoudre un même problème, il est donc difficile de les comparer d'un point de vue théorique vu qu'elles sont basées sur deux approches complètement différentes.

L'objectif de ce chapitre est d'étudier les méthodes existantes de transformation de descripteurs et de proposer une méthode simple pour rendre la distance euclidienne aussi efficace que la distance χ^2. Ainsi, les résultats de classification d'images d'un SVM à noyau RBF basé sur une distance euclidienne devraient être comparables à ceux obtenus par un SVM à noyau RBF basés sur la distance χ^2. Nous présentons, donc, une méthode d'optimisation de descripteurs constituée d'une séquence de transformations élémentaires. Elle permet de réduire le temps de classification en utilisant une distance plus simple à calculer et en autorisant une réduction de dimensions basée sur une ACP. Pour évaluer l'efficacité de la méthode d'optimisation proposée, nous comparons la performance de la classification sur le jeu de données de TRECVid 2010 en utilisant le multi_SVM à noyau RBF [Safa 10] avec la distance euclidienne et la distance χ^2.

L'étape de comparaison a été compliquée à mettre en place à cause des différentes normalisations complémentaires possibles à différents niveaux : au niveau des vecteurs de descripteurs, au niveau des composantes des descripteurs ou au niveau de la combinaison de plusieurs descripteurs. Par ailleurs, nous exposerons une évaluation expérimentale de plusieurs techniques classiques de normalisations de descripteurs : normalisation de longueur (L_1 ou L_2), normalisation min-max (mm), normalisation moyenne nulle et variance unitaire (σ) et la normalisation de puissance.

6.2 Travaux connexes

Les méthodes d'optimisation de descripteurs comprennent un certain nombre de transformations : des normalisations ou des réduction de dimensions. Ces transformations peuvent être appliquées seules ou en séquence.

L'objectif principal des méthodes de normalisation de descripteurs est de les rendre plus invariants à la taille de l'image, à l'éclairage et au contraste, et donc plus facile à comparer entre eux. Ceci se fait en modifiant les valeurs des descripteurs de façon à ce qu'ils adoptent des distributions similaires vis à vis de leur densité ou de leur ordre de grandeur. La normalisation est généralement faite indépendamment au niveau des vecteurs de description (par exemple la normalisation L_1 ou L_2) ou bien au niveau des composantes du vecteur de description (par exemple la normalisation min-max). Néanmoins, d'autres techniques de normalisation traitent directement les éléments des descripteurs sans prendre en considération ni les composantes ni les vecteurs de description (par exemple la transformation de puissance). Les techniques de normalisation les plus populaires sont détaillées dans la section 2.5.

En ce qui concerne la réduction de dimensions, son but principal est de réduire l'espace de stockage et la puissance de calcul nécessaires pour traiter les descripteurs. Les méthodes basées sur l'Analyse de Composantes Principales (ACP) sont les plus répandues [Bish 06]. Ce sont des méthodes statistiques reposant sur des principes de l'algèbre linéaire et qui permettent d'extraire l'information la plus importante à partir d'un ensemble de points (ou vecteurs) pour simplifier ou réduire leurs représentations. Les méthodes de réduction de dimensions ont été détaillées précédemment dans la section 2.5.2.

Suite à la réduction de dimensions par ACP, une deuxième normalisation ou transformation peut être appliquée afin de remettre toutes les composantes du vecteur de description au même niveau. Cette deuxième normalisation peut être une des cinq méthodes de normalisation mentionnées précédemment (L_1, L_2, min-max, variance unitaire ou transformation de puissance) ou une méthode dite de « blanchiment » (ou whitening) [Jego 12a]. La méthode de « blanchiment » consiste à remettre toutes les composantes du vecteur de description au même niveau suite à l'application de l'ACP, le spectre obtenu devenant alors le même que celui d'un bruit blanc. Elle divise chaque composante du vecteur de description par la variance et elle est équivalente à une normalisation de variance unitaire.

6.3 Méthode d'optimisation proposée

La méthode proposée pour l'optimisation de descripteurs est une combinaison de transformations : réduction de dimensions précédée et suivie d'une normalisation comme illustrée dans la figure 6.1. La réduction de dimensions et les différentes normalisations non linéaires ont déjà été étudiées séparément dans la littérature mais

aucun travail s'est intéressé à leur combinaison spécifique ainsi qu'à leur optimisation jointe comme nous le faisons.

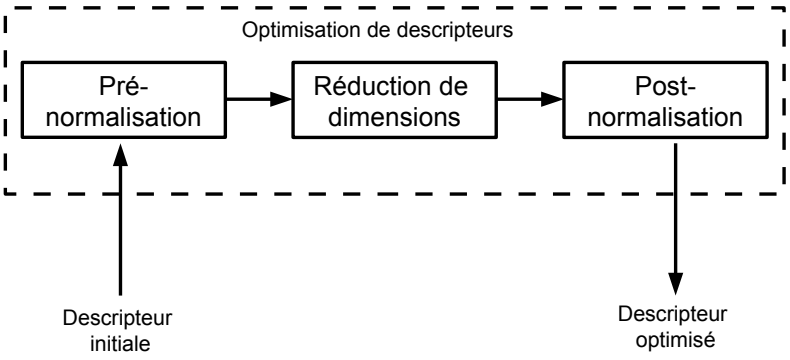


FIGURE 6.1 – La méthode d'optimisation de descripteur proposée

La réduction de dimension améliore la performance des systèmes de classification et de recherche car il a été remarqué que cette réduction de dimensions élimine généralement les composantes redondantes. Comme nous l'avons précisé récemment, les méthodes de normalisation précédant ou suivante la réduction des dimensions (pré- et post-normalisation), peuvent être n'importe quelle combinaison séquentielle des méthodes « élémentaires » citées précédemment : L_1, L_2, unit variance (σ), min-max (mm) ou transformation de puissance (pw).

Les raisons pour lesquelles la pré- et la post-normalisation sont utiles sont différentes. L'application de la transformation avant et après la réduction de dimensions (ACP) ont des effets complètement différents sur les descripteurs. Avant l'ACP leur but est de rendre les descripteurs les plus invariants et les plus équilibrés possibles. Typiquement, la normalisation L_1 ou L_2 rend l'histogramme de description plus invariant à la taille de l'image ou au nombre de points extraits alors que la normalisation à l'échelle de composantes (comme σ ou mm) compense le déséquilibre entre les éléments de l'histogramme. Un objectif supplémentaire de la pré-normalisation est de transformer un descripteur non adapté à la distance euclidienne (comme les descripteurs en histogramme/sacs-de-mots dont la distance la plus adaptée est χ^2) en un descripteur adapté à la distance euclidienne pour une performance similaire ou même meilleure que celle obtenue avec la distance initialement adaptée ; cet objectif est atteint par l'intermédiaire de la transformation de puissance (pw). Après l'ACP, leur but est de réhausser les petites valeurs de composantes par rapport aux plus grandes

valeurs afin de mettre toutes les composantes au même niveau d'importance. Ce but est atteint globalement par une mise à l'échelle linéaire des composantes selon leur variance associée (blanchiment) ou localement par une mise à l'échelle non-linéaire avec une transformation de puissance indépendamment de leur variance associée.

Nos expériences (voir 6.4) montrent que la pré- et la post-normalisation doivent inclure une transformation de puissance, éventuellement combinée avec une autre méthode de normalisation élémentaire. Pour la réduction de dimensions, nous nous sommes contentés de tester celle basée sur l'ACP décrite dans 2.5.2.

Notre processus d'optimisation utilise, donc, pour la pré- et la post-normalisation une transformation de puissance, éventuellement combinable avec une autre méthode de normalisation élémentaire. Dans le cas général, le processus global est contrôlé par trois hyper-paramètres[1] : le nombre k de composantes gardées suite à la réduction de dimensions par l'ACP, α_1 et α_2 les exposants de la transformation de puissance de la pré- et la post-normalisation. Éventuellement, un quatrième hyper-paramètre binaire permet de spécifier l'utilisation ou non d'une autre méthode de normalisation élémentaire en conjonction avec la transformation de puissance.

La première transformation de puissance est appliquée avec l'exposant α_1 qui est optimisé de façon à obtenir les meilleures performances avec la distance euclidienne. La réduction de dimension (ACP) est ensuite appliquée et la valeur de k est optimisée de façon à obtenir la meilleure performance ou bien le meilleur rapport dimension-performance. Enfin, la seconde transformation de puissance est appliquée, avec un α_2 optimisé de façon à obtenir la meilleure performance, sur le résultat de la transformation ACP pour produire le descripteur optimisé final. Toutes les optimisations des hyper-paramètres α_1, α_2 et k sont effectuées par validation croisée sur l'ensemble de développement. Les transformations de notre processus sont séquentielles, il est donc possible d'optimiser les hyper-paramètres de façon séquentielle ou de façon globale. Leur optimisation conjointe pourra être plus coûteuse mais elle pourra entrainer de meilleure performance globale. Lors de nos expérimentations, nous avons tenté d'optimiser une deuxième fois α_1 et k après que les trois paramètres ont été optimisés une première fois mais ceci n'a pas apporté d'améliorations significatives. Cela montre qu'une seule optimisation séquentielle est généralement suffisante.

Pour la post-normalisation, nous avons choisi d'utiliser une seconde fois la transformation de puissance au lieu d'une méthode de « blanchiment ». En effet, les deux méthodes augmentent les petites valeurs par rapport aux grandes valeurs et empêchent les composantes de grande ampleur de dominer et d'éclipser celles avec une faible ampleur. Cependant, le blanchiment s'effectue en se basant sur la variance globale pour une composante donnée alors que la transformation de puissance l'effectue sur chaque élément séparément indépendamment de la variance de la composante à laquelle il appartient. Il est difficile de prédire, en pratique, quelle méthode sera plus efficace et laquelle apportera des améliorations et pour quelles raisons. Toutefois, les expérimentations menées pour comparer ces deux méthodes ont montré

[1] Nous les appelons « hyper-paramètres » car nous pensons qu'ils sont au même niveau que les hyper-paramètres d'un classificateur, par exemple C et γ dans les SVMs à noyau RBF.

que toutes les deux apporteraient des améliorations mais que la transformation de puissance est généralement légèrement plus efficace que la méthode de blanchiment (voir section 6.4).

6.4 Évaluations

Les expérimentations sur l'optimisation de descripteurs ont été conduites sur le jeu de données de TRECVid 2010 pour la tâche d'indexation sémantique (SIN). Le jeu de données est composé de deux grands ensembles de documents vidéos : l'ensemble d'apprentissage et l'ensemble de test. L'ensemble d'apprentissage contient 119 685 plans de 3 173 vidéos avec une moyenne de 37 plans par vidéo alors que l'ensemble de test est composé de 146 788 plans de 8 467 vidéos avec une moyenne de 17 plans par vidéo. Les plans vidéos sont les échantillons dans lesquels les concepts sont recherchés.

6.4.1 Descripteurs vidéos

Nous avons utilisé plusieurs descripteurs de différents types et de différentes tailles, qui ont été produits et partagés par les nombreux partenaires du projet IRIM de GDR-ISIS [Gori 10]. La plupart des descripteurs choisis sont basés sur des histogrammes de couleurs ou sur des approches d'agrégation en sacs-de-mots. Nous comparons, toutefois, les méthodes d'optimisation sur différents types de descripteurs, comme ceux basés sur les filtres Gabor pour représenter la texture ou ceux modélisant le contenu audio. Au final, 12 descripteurs ont été sélectionnés :

- **lab1×3×192 et qwm1×3×192 :** descripteurs basés sur la concaténation d'histogrammes [Gori 10], le premier utilise les couleurs CIE LAB alors que le deuxième utilise les ondelettes quaternioniques (3 échelles et 3 orientations). Les histogrammes sont calculés pour 3 parties verticales et la taille du dictionnaire est égale à 192. Les deux descripteurs comportent 576 dimensions.

- **sm462 :** descripteur de Moments de Salliance (SM) [Redi 11] ; c'est un descripteur global qui intègre de l'information analysée localement. Le descripteur résultant comporte 462 dimensions.

- **audioSpectro :** descripteur audio représentant le profil spectral en 28 bandes sur une échelle Mel, normalisé et à 28 dimensions.

- **dense_sift_k512 :** sacs-de-SIFT calculé sur un histogramme, à 512 dimensions.

- **h3d64 :** un histogramme RGB normalisé $4 \times 4 \times 4$, à 64 dimensions.

- **gab40 :** un descripteur normalisé basé sur une transformation de Gabor, 8 orientations $\times$ 5 échelles, à 40 dimensions.

- **hg104** : une fusion précoce (concaténation) de h3d64 et gab40, à 104 dimensions.

- **sift_<méthode>_[unc] :** sacs-de-mots visuels, opponent SIFT, généré par le programme de Koen van de Sande [Van 10]. <méthode> est relié à la façon par laquelle les points SFIT sont sélectionnés : **har** correspond à un filtrage via le détecteur Harris-Laplace et **dense** correspond à un échantillonnage dense ; la version avec **_unc** correspond au même descripteur avec du flou introduit dans le calcul de l'histogramme. Nous avons utilisé quatre descripteurs de ce type. La taille du dictionnaire, et donc du descripteur, est égale à 1000.

6.4.2 Optimisation des paramètres

Pour évaluer chaque méthode de normalisation, nous avons utilisé une approche multi-apprentissage basée sur un classificateur SVM à noyau RBF (MSVM) [Safa 10]. MSVM est un ensemble de méthode d'apprentissage basé sur un SVM « standard » pour gérer les problèmes de grand déséquilibre dans les données. Les paramètres à optimiser sont les hyper-paramètres γ du noyau RBF du classificateur SVM et α de la transformation de puissance. L'optimisation est faite par validation croisée et selon la métrique la moyenne de la précision moyenne (ou MAP) sur les 30 concepts de l'ensemble d'apprentissage de TRECVid 2010. Dans ce qui suit, nous présentons le processus d'optimisation et une comparaison entre les différentes méthodes de normalisation.

6.4.3 Évaluation des méthodes de normalisation de référence

Toutes les méthodes de normalisation de référence sont des méthodes sans paramètres à fixer. Les deux méthodes de normalisation niveau descripteur L_1 et L_2 ainsi que les deux méthodes de normalisation niveau composantes min-max (mm) et la variance unitaire (σ) peuvent être évaluées séparément ou combinées.

Les tableaux 6.1 et 6.2 montrent la performance des systèmes sur l'ensemble d'apprentissage de TRECVid 2010, avec la distance euclidienne et la distance χ^2 respectivement, en utilisant les méthodes de normalisation de référence et quelques combinaisons de certaines d'entre elles (d'autres combinaisons ont été essayées mais elles ont été moins efficaces). Les résultats concernant la normalisation L_1 ne sont pas affichés dans les tableaux car ils ont été très proches de ceux obtenus avec la normalisation L_2. Les résultats obtenus après la normalisation sont comparés avec ceux obtenus avec la même méthode d'apprentissage utilisant les deux distances mais sans aucune normalisation (raw). Comme nous pouvons le voir dans ces tableaux, la performance du système varie significativement avec les différentes normalisations. Pour la normalisation L_2, σ et mm, la performance est souvent très proche de celle de la méthode (raw) et la meilleure normalisation entre elles n'est pas la même pour tous les descripteurs considérés. Comme attendu, la distance χ^2 est plus efficace que la distance euclidienne pour les descripteurs sous la forme d'histogramme (les 8

derniers descripteurs) mais elle ne fournit pas de différence significative pour les descripteurs qui ne sont pas sous la forme d'histogramme (les 4 premiers descripteurs).

Descripteur	Raw	L_2	σ	mm	L_2-σ	σ-L_2	L_2-mm	mm-L_2
sm462	0.95	1.21	**1.89**	1.15	**3.17**	2.35	1.52	1.15
audioSpectro	1.55	1.56	1.38	**1.57**	1.48	1.33	**1.58**	1.54
gab40	**2.65**	2.57	2.40	1.82	**2.67**	2.50	1.95	1.42
hg104	3.68	3.66	**4.07**	2.78	4.08	**4.21**	3.23	2.84
h3d64	1.58	1.59	**2.55**	1.61	2.27	**2.48**	1.52	1.60
labm1x3x192	3.46	3.42	3.16	**3.55**	3.26	3.46	3.48	**3.59**
qwm1x3x192	3.12	3.51	3.56	**3.73**	3.76	**4.69**	3.62	4.37
dense_sift_k512	5.72	6.10	**6.95**	6.36	6.84	**7.33**	6.76	6.66
sift_har	5.07	**5.29**	4.85	4.55	4.70	4.72	4.69	**5.00**
sift_har_unc	5.39	**5.40**	5.10	5.16	5.14	5.04	5.13	**5.17**
sift_dense	4.41	4.49	**5.45**	4.94	4.99	**5.59**	5.11	5.07
sift_dense_unc	4.46	4.72	**6.17**	5.91	5.34	**6.26**	5.48	5.99

TABLE 6.1 – La MAP (en pourcentage) obtenue avec les méthodes de normalisation de référence utilisant la distance euclidienne sur les données d'apprentissage de TRECVid 2010.

Descripteur	Raw	L_2	σ	mm	L_2-σ	σ-L_2	L_2-mm	mm-L_2
sm462	1.44	1.55	**2.43**	1.49	**3.15**	2.19	1.92	1.36
audioSpectro	0.30	0.19	0.17	**0.96**	0.31	0.33	**1.50**	1.25
gab40	**2.47**	2.15	2.40	1.86	**2.44**	2.38	1.92	1.49
hg104	3.78	3.87	**4.47**	.50	4.67	**4.77**	3.63	3.33
h3d64	0.81	1.24	**1.37**	1.12	**3.26**	2.99	2.27	2.54
labm1x3x192	4.24	3.99	3.79	**4.35**	3.90	3.80	**4.23**	3.94
qwm1x3x192	**5.04**	4.55	4.17	4.30	4.15	**4.91**	4.39	4.76
dense_sift_k512	7.84	7.60	7.62	**8.41**	7.73	8.14	**8.20**	7.98
sift_har	**4.16**	3.70	3.67	3.34	4.60	4.48	**4.67**	4.66
sift_har_unc	**4.85**	4.53	4.25	4.32	**5.13**	4.99	5.11	5.12
sift_dense	6.23	**6.26**	5.86	5.72	5.26	**5.63**	5.37	5.46
sift_dense_unc	6.99	**7.46**	6.88	6.76	5.73	**6.52**	5.80	6.14

TABLE 6.2 – La MAP (en pourcentage) obtenue avec les méthodes de normalisation de référence utilisant la distance χ^2 sur les données d'apprentissage de TRECVid 2010.

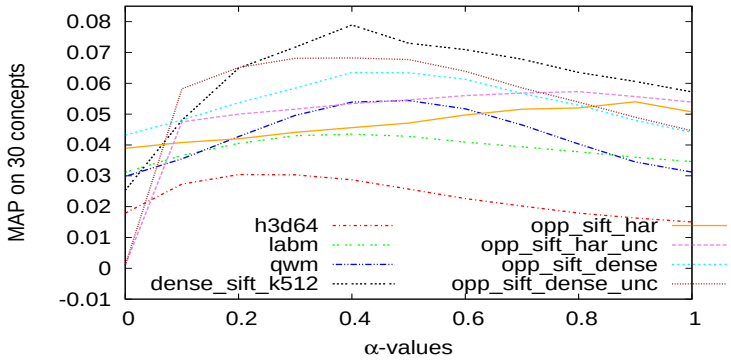
6.4.4 Évaluation de la transformation de puissance

La transformation de puissance possède un seul paramètre à optimiser, l'hyperparamètre α. Pour comparer les méthodes de normalisation, nous avons d'abord besoin de trouver la valeur optimale de α pour chaque descripteur. Nous avons cherché

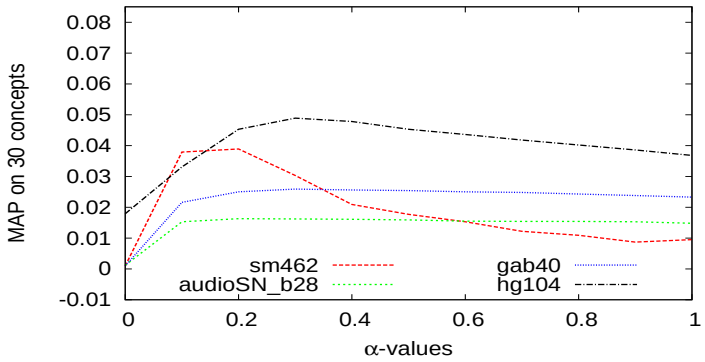
la valeur optimale par descripteur parmi 10 différentes valeurs de $\alpha \in [0, 1]$. La transformation de puissance a été évalué en coordination avec la meilleure combinaison de méthodes normalisation de référence (trouvée dans 6.4.3) pour chaque descripteur.

Les figures 6.2 et 6.3 montrent les résultats de l'optimisation de α pour les deux distances considérées : euclidienne et χ^2, respectivement. Chaque courbe dessine la performance du système (en MAP) en fonction des différentes valeurs de l'exposant α pour un descripteur donné. Comme nous pouvons le constater le paramètre α possède plusieurs valeurs optimales pour chaque descripteur et chaque distance. Ceci montre l'importance du choix de la meilleure valeur de α. Par exemple, avec la distance euclidienne le descripteur h3d64 obtient de meilleure performance avec un $\alpha = 0.3$, le descripteur dense_sift_k512 obtient la meilleure performance avec $\alpha = 0.4$.

Il est intéressant de noter que la valeur optimale de α avec la distance χ^2 correspond approximativement au double de celle de la distance euclidienne. La valeur optimale pour la distance euclidienne est souvent proche de 0.5.

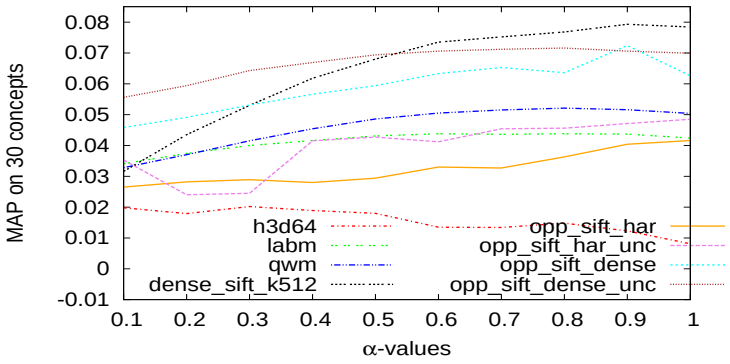


(a) Descripteurs ayant la forme d'histogramme / BoW

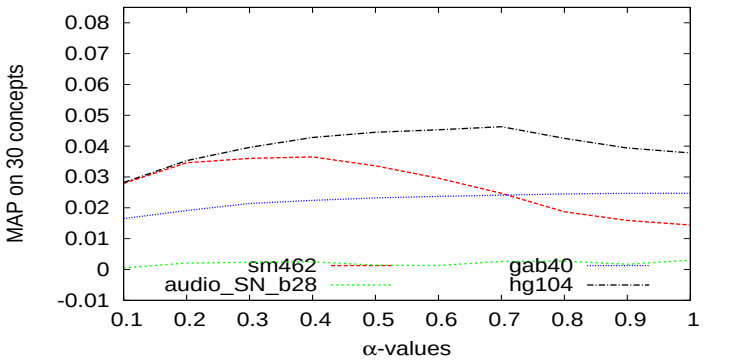


(b) Descripteurs n'ayant pas la forme d'histogramme

FIGURE 6.2 – Réglages de l'hyper-paramètre de la transformation de puissance α utilisant la distance euclidienne sur les données d'apprentissage de TRECVid 2010.



(a) Descripteurs ayant la forme d'histogramme / BoW



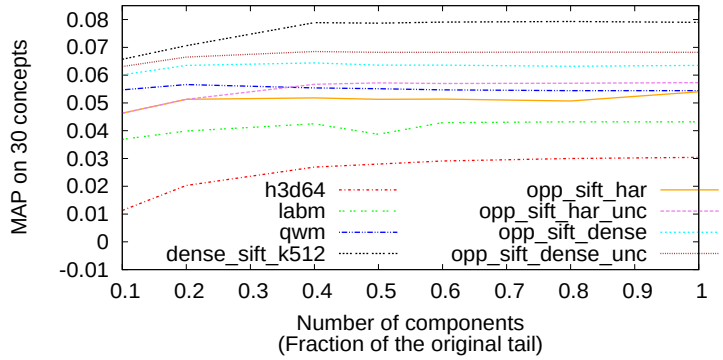
(b) Descripteurs n'ayant pas la forme d'histogramme

FIGURE 6.3 – Réglages de l'hyper-paramètre de la transformation de puissance α utilisant la distance χ^2 sur les données d'apprentissage de TRECVid 2010.

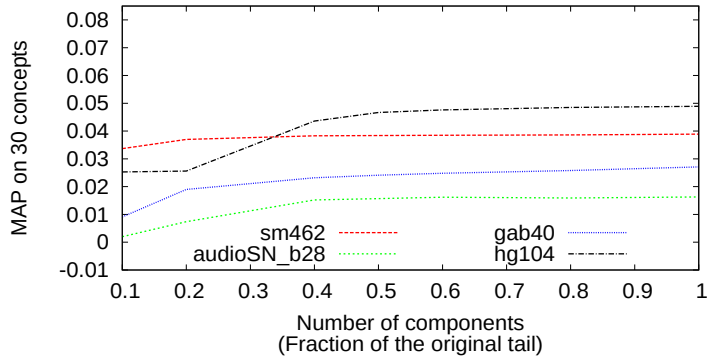
6.4.5 Évaluation de la réduction de dimensions par ACP

La figure 6.4 montre la performance du système (en MAP) obtenue en appliquant la transformation de puissance suivie d'une réduction de dimensions basée sur l'ACP. Pour les descripteurs de petite taille, l'application de l'ACP n'est pas utile. En revanche, notre objectif principal est de montrer l'impact de l'utilisation d'une réduction de dimension basée sur l'ACP sur les descripteurs à grandes dimensions sur la performance des systèmes de classification. Nous avons testé plusieurs valeurs de l'hyper-paramètre k de l'ACP (le nombre de composantes importantes) sur chacun des descripteurs considérés, ces valeurs correspondent à une fraction allant de 0.1 à 1 du nombre de dimension original.

Comme nous pouvons voir dans la figure 6.4, le nombre de composantes optimal varie en fonction du descripteur. Pour les descripteurs de grandes tailles, nous avons fixé la valeur optimale de k à la plus petite valeur donnant la performance la plus élevée ou presque. Par exemple, la valeur choisie pour k pour le meilleur descripteur (dense_sift_k512) est de $0.4 * 512 = 204$.



(a) Descripteurs ayant la forme d'histogramme / BoW



(b) Descripteurs n'ayant pas la forme d'histogramme

FIGURE 6.4 – Évaluation de l'ACP avec la distance euclidienne sur les données d'apprentissage de TRECVid 2010. Les courbes montrent seulement les résultats avec la distance euclidienne pour les descripteurs sous la forme d'histogramme (a) ou non (b).

6.4.6 Évaluation de la transformation de puissance avec une réduction de dimensions par ACP et une transformation post-ACP

Alors que les expérimentations précédentes concernaient l'analyse des paramètres pertinents sur l'ensemble d'apprentissage, nous abordons ici l'évaluation de la méthode proposée et nous exposons les résultats obtenus par notre méthode sur l'ensemble des données de test de TRECVid 2010.

Nous avons évalué différentes combinaisons de méthodes de transformations sur l'ensemble de test de TRECVid 2010 après avoir optimisé tous les paramètres reliés par validation croisée sur l'ensemble de développement, pour chaque descripteur et combinaison de méthodes. Les combinaison de méthodes testée ou les sujets de tests comprennent :

- la meilleure combinaison de méthodes de normalisation de références selon 6.4.3 (baseline) ;
- la même avec une transformation de puissance pré-ACP ;
- la précédente suivie d'une réduction de dimension d'ACP ;
- la précédente avec en plus une transformation de puissance post-ACP ou un blanchiment post-ACP.

Les résultats sont présentés pour la distance euclidienne et la distance du χ^2 pour les deux premières méthodes et que pour la distance euclidienne pour les reste (vu que la distance χ^2 devient insensée après l'application d'une ACP).

Selon les sujets de test, les hyper-paramètres suivants ont été optimisés : α_1 l'exposant de la transformation de puissance pré-ACP ; k le nombre optimal de composantes à garder après ACP ; α_2 l'exposant de la transformation de puissance post-ACP ; et $(B; \beta)$ les paramètres de blanchiment optimaux. Le tableau 6.3 contient les valeurs optimales trouvées pour ces paramètres pour chacun des descripteurs considérés. d est la dimension originale d'un descripteur.

Les résultats pour les 12 descripteurs considérés sont affichés dans le tableau 6.4. Ils sont cohérents avec ceux obtenus sur les données d'apprentissage, ce qui prouve la robustesse et la bonne capacité de généralisation de la méthode proposée. Le tableau 6.4 révéle l'efficacité de la transformation de puissance (+pw) avec les deux distances. Il montre également l'efficacité de la réduction de dimensions ACP avec la distance euclidienne et la performance des deux normalisations après l'ACP (+pca) : une seconde transformation de puissance (+pw) et un blanchiment (+wh).

La transformation de puissance donne de meilleurs résultats que toutes les autres méthodes de normalisation évaluées pour tous les descripteurs considérés. Elle est même meilleure avec la distance euclidienne qu'avec la distance χ^2 dans la grande majorité des cas. L'utilisation de la réduction de dimensions par ACP rend le système plus rapide tout en préservant ou même augmentant les performances du système. La deuxième transformation de puissance améliore la performance plus que le blanchiment dans la plupart des cas. Enfin, la combinaison proposée « Euc+pw+pca+pw » est

Descripteur	d	α_1	k (σ_k^2/σ_d^2)	α_2	(B, β)
sm462	462	0.2	277 (0.999)	0.6	(4.0, 0.4)
audioSpectro	28	0.2	28 (1.000)	0.8	(2.5, 0.5)
gab40	40	0.3	40 (1.000)	0.6	(8.0, 0.6)
hg104	104	0.3	104 (1.000)	0.6	(8.0, 0.6)
h3d64	64	0.3	64 (1.000)	0.6	(4.0, 0.6)
labm1x3x192	576	0.4	346 (0.980)	0.5	(4.0, 0.7)
qwm1x3x192	576	0.5	115 (0.931)	0.7	(2.0, 0.1)
dense_sift_k	512	0.4	204 (0.931)	0.8	(2.0, 0.4)
sift_har	1000	0.9	400 (0.734)	0.7	(2.0, 0.9)
sift_har_unc	1000	0.8	500 (0.832)	0.8	(2.0, 0.9)
sift_dense	1000	0.4	400 (0.827)	0.8	(2.0, 0.9)
sift_dense_unc	1000	0.4	400 (0.933)	0.8	(2.5, 0.4)

TABLE 6.3 – Les valeurs optimales des paramètres de la méthode de normalisation proposée par descripteur avec la distance euclidienne.

souvent la meilleure méthode d'optimisation et, dans les cas où elle ne l'est pas, elle est très proche de la meilleure. Ceci a été vérifié même avec l'application d'une fusion ou d'un post-traitement comme mentionné ci-dessous.

Les résultats ont été aussi présentés pour un système qui effectue une simple fusion tardive (moyenne des scores de classification). La fusion a été essayée séparément uniquement pour les descripteurs étant sous la forme d'histogramme (fusion8), ceux ne l'étant pas (fusion4) et pour tous les descripteurs (fusion-All). La transformation de puissance a obtenu les meilleures performances avec une fusion qui a atteint un score de 7.07% pour la fusion de tous les descripteurs (fusion-All) avec une ACP et une distance euclidienne. Elle atteint 8.07% après une étape de post-traitement : un reclassement basé sur le contexte temporel [Safa 11b] (cette étape de post-traitement exploite l'homogénéité statistique globale et locale du contenu vidéo).

La MAP globale obtenue de 8.07% peut être comparé avec le résultat obtenu par le meilleur système participant à TRECVid 2010 (SIN) de 9.00% de MAP. Sachant que plus de descripteurs peuvent être utilisés ; que la fusion appliquée est basique ; et qu'un post-traitement plus avancé des classificateurs fusionnés peut encore améliorer les performances. Par exemple un re-classement basé sur le contexte conceptuel [Hama 12] pourrait être utilisé.

6.4.7 Temps d'exécution

Toutes les expériences ont été menées sur des machines avec deux processeurs quadruple cœur cadencés à 2.66 Ghz et avec 32 Go de mémoire vive. Le temps d'exécution dépend de la taille du descripteur. Le temps d'apprentissage et d'indexation cumulé a été mesuré pour les 30 concepts et pour les 12 descripteurs considérés. Nous reportons dans le tableau 6.5 le temps d'exécution total du processus de classification (d'apprentissage + d'indexation) en nombre d'heures de traitement pour les 30 concepts.

Descripteur	Euc (baseline)	χ^2 (baseline)	Euc +pw	χ^2 +pw	Euc +pw +pca	Euc +pw +pca +pw	Euc +pw +pca +wh
sm462	1.04	0.57	2.46	1.78	2.33	**3.39**	2.73
audioSpectro	0.07	0.07	0.11	0.06	0.35	**0.36**	0.33
gab40	1.06	1.03	1.15	1.04	1.14	**1.46**	1.44
hg104	1.77	2.14	2.46	2.07	2.40	2.76	**2.85**
Fusion4	2.95	3.07	3.66	3.16	4.03	**4.99**	4.61
h3d64	0.53	0.54	**1.45**	0.46	1.26	1.40	1.27
labm1x3x192	1.26	2.38	2.70	**2.88**	2.65	2.75	2.55
qwm1x3x192	1.42	2.13	2.17	2.27	2.14	**2.41**	2.13
dense_sift_k512	3.89	4.20	4.18	3.77	4.05	**4.56**	4.44
sift_har	2.28	1.87	2.23	1.54	2.49	**2.54**	2.33
sift_har_unc	2.68	2.84	2.93	2.60	**3.13**	3.10	3.09
sift_dense	3.32	3.40	3.75	3.46	3.81	**3.99**	3.76
sift_dense_unc	3.81	4.33	4.33	4.51	4.26	**4.57**	4.34
Fusion8	4.76	6.00	6.24	6.10	6.25	**6.69**	6.51
Fusion-All	5.23	6.04	6.32	6.18	6.46	**7.07**	6.88
Re-ranking	6.24	6.74	7.23	6.83	7.31	**8.07**	7.63

TABLE 6.4 – La MAP (en pourcentage) sur les données de test de TRECVid 2010, en utilisant les différentes méthodes de normalisation avec la distance euclidienne (Euc) ou la distance χ^2.

Le temps d'exécution est donné pour les deux distances (euclidienne et χ^2) sans ACP et pour la distance euclidienne avec ACP, toutes avec la transformation de puissance optimale. Le temps d'exécution total est de 201 heures pour la distance optimale de χ^2, de 110 heures avec la distance euclidienne et de 53 heures avec la réduction de dimension par ACP. Comme nous pouvons noter, le MSVM-RBF avec la distance euclidienne est significativement plus rapide que la version originale avec χ^2. De même, après l'application de l'ACP, le système est bien plus rapide pour une performance équivalente ou même améliorée.

	χ^2	Euc	PCA-Euc
sm462	17.2	9.8	5.5
audioSpectro	1.5	1.4	0.6
gab40	2.2	1.2	1.1
hg104	3.6	2.4	1.9
h3d64	1.5	0.9	0.8
labm1x3x192	17.6	18.8	6.1
qwm1x3x192	15.8	15.8	3.6
dense_sift_k512	21.6	13.2	3.3
sift_har	15.5	10.4	8.1
sift_har_unc	35.2	12.6	8.5
sift_dense	24.9	11.4	7.8
sift_dense_unc	44.5	12.0	5.7
Total	201.1	109.9	53.0

TABLE 6.5 – Le temps d'exécution (en nombre d'heures sur 8 cœurs) pour les 30 concepts des données de test de TRECVid 2010.

6.4.8 Applications sur les descripteurs à très grandes dimensions

La méthode proposée a également été appliquée à grande échelle dans le contexte de TRECVid 2012 à la tâche d'indexation sémantique. La collection contient 545 923 plans vidéos dont 400 289 pour l'ensemble d'apprentissage et 145 634 pour l'ensemble de test, et 346 concepts à classifier. IRIM produit une dizaine de différents types de descripteurs, un grand nombre d'entre eux possède plusieurs variantes (comme la taille du dictionnaire pour les sacs-de-mots) générant au final plus de 100 descripteurs de qualité moyenne à très bonne [Ball 12]. La dimension de ces descripteurs varie entre 15 et 32 768. Notre approche a été appliquée à la plupart d'entre eux. Pour les descripteurs de grandes dimensions, la valeur de optimale trouvée de k a toujours été bien inférieure à la taille originale d des descripteurs et n'a jamais excédé 768 même pour un $d \geq 10K$. Ceci révèle que ces descripteur de grandes dimensions sont extrêmement redondants. Cette réduction de dimensions a encore été obtenue avec une augmentation simultanée de la performance de la classification.

Le tableau 6.6 donne une idée rapide des résultats d'application de la méthode proposée à certains descripteurs de grandes dimensions à TRECVid 2012 à l'indexation sémantique. La tendance observée sur les résultats obtenus sur les données de

TRECVid 2010 a été confirmée sur TRECVid 2012 avec très peu d'exceptions, la seule notable étant celle d'un type de descripteurs, les vecteurs de tenseurs localement agrégés (ou VLAT). Ces derniers n'ont pas tiré bénéfice de l'ensemble des transformations de notre méthode. Cela est probablement dû au fait que ces descripteurs intègrent déjà une forme d'ACP. Cependant, notre approche fonctionne bien sur les descripteurs basés sur les noyaux de Fisher.

Descripteur	d	α_1	UL	k	α_2	MAP
CEALIST/bov_dsiftSC_8192	8192	0.80	-	256	0.70	17.74
CEALIST/bov_dsiftSC_21504	21504	0.70	-	512	0.80	19.67
ETIS/labm2x2x1024	4096	0.35	-	512	0.80	11.31
ETIS/qwm2x2x1024	4096	0.45	-	512	0.80	11.24
INRIA/dense_sift_k1024	1024	0.45	-	256	0.50	15.41
INRIA/dense_sift_k2048	2048	0.45	-	320	0.60	16.98
INRIA/dense_sift_k4096	4096	0.45	-	400	0.70	18.08
INRIA/dense_sift_k8192	8192	0.45	-	512	0.70	18.63
INRIA/vlad_10240	10240	0.50	L_2	640	0.60	20.72
INRIA/vlad_20480	20480	0.50	L_2	640	0.60	21.03
INRIA/vlad_32768	32768	0.50	L_2	640	0.60	18.07
LIF/percepts_5_3_1_15	225	0.60	-	225	0.40	11.19
LIF/percepts_10_6_1_15	900	0.50	-	256	0.40	11.29
LIF/percepts_20_13_1_15	3900	0.50	-	256	0.50	11.67
LIRIS/OCLBP_DS_4096	4096	0.60	L_2	512	0.90	6.88
LSIS/mlhmslbp_spyr_10240	10240	0.50	-	768	0.35	15.51
LSIS/mlhmslbp_spyr_26624	26624	0.70	-	768	0.35	14.67

TABLE 6.6 – La méthode proposée appliquée sur quelques descripteurs à grandes dimensions à la tâche d'indexation sémantique de TRECVid 2012. d et k sont les dimensions des descripteurs avant et après l'application de la réduction de dimensions par ACP ; IL indique si une normalisation de longueur (L_1 ou L_2) a été appliquée ou pas ; MAP (en pourcentage) est la performance du descripteur optimisé uniquement.

La performance globale du système après une fusion tardive optimisée et combinée avec un re-classement selon le contexte temporel et conceptuel est de 26.92% pour notre meilleure soumission alors que la performance du meilleur système participant à TRECVid SIN 2012 est de 32.20% sachant que ce système utilise des annotations supplémentaires non officielles.

6.5 Conclusion

Nous avons proposé et évalué une méthode pour optimiser les descripteurs utilisés pour la recherche et l'indexation des contenus multimédias. La méthode proposée combine une réduction de dimension basée sur une ACP avec des transformations non linéaires avant et après l'ACP. La transformation résultante est globalement optimisée. Les descripteurs produits possèdent beaucoup moins de dimensions alors

qu'ils fournissent le plus souvent de meilleures performances avec la distance euclidienne que les descripteurs originaux avec leur distance optimale, généralement χ^2, plus coûteuse à calculer. Ils donnent aussi de meilleurs résultats que les mêmes descripteurs optimisés avec des méthodes de normalisation classiques comme L_1, L_2 ou à l'échelle de la composante (min-max), de la normalisation de variance ou de leurs simples combinaisons.

La méthode a été validée et évaluée sur plusieurs descripteurs avec les données de la tâche d'indexation sémantique de TRECVid 2010. Elle a ensuite été utilisée à grande échelle sur la tâche d'indexation sémantique de TRECVid 2012, TRECVid 2013 et TRECVid 2014, sur des dizaines de descripteurs de différents types et dont les dimensions originales varient de 15 à 32 768. La même transformation peut être utilisée également pour la recherche multimédia dans le contexte d'exemples requêtes et/ou de retour de pertinence.

7 Conclusion et perspectives

Pour conclure notre travail, nous récapitulons dans ce chapitre nos principales contributions puis nous exposerons les perspectives générales ouvertes par ces travaux.

7.1 Synthèse et contributions

Dans cette thèse, nous nous sommes intéressés à l'indexation automatique par le contenu des documents multimédias. Nous avons exploré et présenté différentes pistes pour améliorer la performance de ces systèmes pour faire face à la croissance continue de la quantité des documents multimédia (images et vidéos). Dans un premier temps, nous avons abordé les méthodes de fusion en proposant une nouvelle méthode de fusion inter-modalité « doublement précoce » pour mieux exploiter les corrélations entre les différentes modalités. Nous nous sommes ensuite penchés la problématique de localisation des concepts basiques (comme des objets) dans les images. Nous nous sommes également attaqués au problème d'annotations inexactes des données d'apprentissage. Enfin, nous avons traité la problématique d'optimisation des descripteurs en réduisant leur taille tout en augmentant leur performance. Nous avons intégré nos différentes contributions au processus d'indexation sémantique de pointe décrit dans la section 2.3 et nous les avons évalué séparément sur des données complexes de TRECVid ou de MediaEval.

Notre première contribution s'inscrit dans le cadre de la fusion de différentes modalités ou sources d'information. Comme nous avons montré dans l'état de l'art, une multitude de descripteurs est disponible, actuellement, pour représenter les différentes modalités des documents multimédias (visuelles, audio, mouvement, ...). De plus l'utilisation de plusieurs descripteurs augmente la performance des systèmes d'indexation, nous avons abordé la problématique de fusion entre modalités et leur capacité à capturer le maximum d'information conjointe à deux modalités. Notre

avons proposé une nouvelle méthode de fusion dite « doublement précoce » pour représenter conjointement le contenu audio-visuel d'une vidéo. Elle exploite la corrélation entre l'information audio et l'information visuelle en construisant un dictionnaire audio-visuel joint dans le but de découvrir des motifs spécifiques audio-visuels. La fusion « doublement précoce » nous a permis d'obtenir les meilleurs résultats dans le cadre de la détection de scènes violentes à MediaEval 2013 (premier pour la violence objective et deuxième pour la violence subjective). De plus, nous pensons que ces travaux ouvrent de nouvelle direction de recherche dans le domaine de la fusion multi-modale.

Notre deuxième contribution concerne la localisation d'objets dans les vidéos difficiles et réelles en opposition à celles qui sont encadrées et jouées. Nous avons proposé une méthode faiblement supervisée qui tente de détecter de l'invariabilité spécifique à un concept donné dans la variabilité globale d'une vidéo. Notre méthode consiste à créer un nouveau modèle discriminant basé sur l'occurrence statistique de descripteurs locaux invariants à partir de données d'entraînement faiblement annotées (le lieu d'apparition des objets dans les images n'est pas connu).
Lors des évaluations, nous avons manqué d'éléments de comparaison mais la méthode a montré des résultats encourageants pour la suite des travaux sur la localisation faiblement supervisée.

Notre troisième contribution porte sur la réduction du bruit dans les annotations au niveau des plans dans le contexte d'un apprentissage supervisé pour la détection de concepts ou d'événements. En effet, la majorité des corpus de vidéos disponibles actuellement sont annotées uniquement au niveau de la vidéo entière à cause des coûts élevées des annotations manuelles. Or l'hétérogénéité du contenu des vidéos entraine une incohérence entre le contenu visuel de certains plans et l'étiquette attribuée à la vidéo. Dans ce but, nous avons proposé deux méthodes. La première méthode génère de nouvelles annotations au niveau des plans à partir d'annotations fournies au niveau de vidéos complètes. Cette méthode cherche à trier l'ensemble des plans des vidéos positives pour enlever un maximum de plans négatifs (faux positifs) et garder un maximum de vrais positifs. La deuxième méthode, quant à elle, pondère les plans des vidéos d'apprentissage pour limiter le bruit. Elle cherche à reconnaître les plans vrais positifs dans l'ensemble des plans des vidéos positives et à leur attribuer des poids plus importants dans le contexte d'un apprentissage supervisé pondéré pour la détection de concepts ou d'événements. Ces deux idées, théoriquement intéressantes à priori, n'ont pas permis de concevoir un algorithme capable d'améliorer la performance de l'état de l'art. Les deux méthodes se sont avérées non adaptées aux données d'apprentissage comportant un grand déséquilibre entre les classes. Cependant nous pensons que c'était une piste intéressante à explorer.

Notre deuxième et troisième contribution sont reliées entre elles vu qu'elle relèvent toutes les deux du problème de l'Apprentissage d'Instances Multiples.

Enfin, pour notre quatrième contribution, nous nous sommes penchés sur la qua-

lité des descripteurs représentant le contenu d'une image (ou d'une vidéo). Ces descripteurs ont de plus en plus tendance à être de grande dimension ce qui rend leur traitement difficile sur les grands corpus. Notre quatrième contribution est une méthode pour optimiser les descripteurs utilisés en trouvant un compromis entre leur dimension et leur capacité à représenter le contenu. Elle combine une réduction de dimension basée sur une ACP avec des transformations non linéaires avant et après l'ACP. Nous avons montré expérimentalement l'utilité de notre optimisation, les descripteurs optimisés avec notre méthode possèdent beaucoup moins de dimensions alors qu'ils fournissent le plus souvent de meilleures performances que les descripteurs originaux optimisés avec des méthodes de normalisation classiques (comme L_1, L_2, ...). Notre méthode d'optimisation est systématiquement intégrée au système d'indexation évalué annuellement à la campagne d'évaluation TRECVid et elle a participé au bon classement (troisième et deuxième places) obtenu en 2012 et 2013.

7.2 Perspectives

Plusieurs améliorations peuvent être apportées aux différentes méthodes proposées et de nombreuses perspectives sont alors envisageables. Dans ce qui suit nous nous focalisons sur les perspectives les plus intéressantes.

L'utilisation de la fusion jointe audio-visuelle a été effectué au niveau d'un plan. Cependant, la corrélation entre les modalités peut être mieux capturée avec une localisation temporelle plus étroite qu'un plan entier. Par conséquent, dans le futur l'utilisation de la fusion à une échelle plus petite que celle du plan peut être envisagée. D'une autre part, notre fusion doublement précoce a uniquement été évalué sur deux descripteurs : MFCC pour l'audio et STIP-HOF pour le mouvement. Notre travail pourrait être étendu pour intégrer de nouveaux types de descripteurs et fusionner potentiellement plus que deux descripteurs. Enfin, une autre piste possible est celle de l'application de cette représentation conjointe à d'autres types de concepts dynamiques que la violence sur laquelle nous avons évalué la méthode de fusion.

En ce qui concerne la localisation d'objets dans les images, notre modèle dessine le cadre englobant autour de l'objet selon une méthode simple calculant les histogrammes de la projection horizontale et verticale des points de l'image. Dans le futur, d'autres techniques plus avancées pourront être utilisées pour le dessin du cadre englobant afin d'améliorer la localisation au lieu de se servir de simple rectangles. De plus, le modèle actuel ne prend pas en compte le cas où l'objet apparaît plusieurs fois dans la même image, nous pourrons compléter notre modèle pour gérer ce cas de figure. Par ailleurs, la performance du système peut être augmentée en utilisant d'autres types de descripteurs ou même une fusion de plusieurs descripteurs au lieu de l'utilisation unique des SIFT.

Les deux méthodes proposées pour le filtrage du bruit causé par des annotations

inexactes utilisent le contenu visuel des plans et des vidéos. Elles se basent sur l'idée que les plans contenant un concept ou un événement sont semblables alors que les plans ne représentant pas ce concept ou cet événement sont différentes entre eux et du reste des plans. Notre étude s'est limitée à l'utilisation de la distance moyenne et de la densité pour la mise en pratique de l'idée mais d'autres formules peuvent être considérées. De plus, nous avons utilisé uniquement un descripteur de couleur et de texture alors qu'une multitude d'autres descripteurs pourront être utilisés.

Notre méthode d'optimisation a permis d'améliorer la capacité des descripteurs à représenter le contenu multimédia et à améliorer la performance des systèmes d'indexation. En revanche, seulement la fusion tardive a été testée et séparément de la transformation des descripteurs. Dans les travaux futurs, nous pourrons considérer une combinaison de notre approche avec une fusion précoce de plusieurs descripteurs.

Enfin, d'un point de vue général nous prévoyons d'explorer les méthodes actuelles bio-inspirées pour l'indexation des vidéos. Ces méthodes s'appuient sur le fonctionnement du cerveau ou du système de vision humaine. Certaines méthodes exploitent des propriétés du système visuel humain pour améliorer les descripteurs, comme le descripteur FREAK [Alah 12] ou Retina-SIFT qui se concentrent sur les propriétés de la rétine humaine pour rendre les descripteurs plus robustes aux dégradations des images et plus sensibles aux informations spatiales et temporelles [Stra 14]. D'autres méthodes se basent sur le fonctionnement hiérarchique du cerveau humain pour proposer une alternative aux processus de classification supervisée classiques et effectuer un apprentissage profond (*Deep Learning*) pour résoudre le problème de variations des représentations [Kriz 12]. En observant les résultats obtenus par ces méthodes bio-inspirées nous constatons leurs potentiels et toutes les pistes d'améliorations encore possibles et surtout dans le cadre de l'indexation des vidéos. Une piste intéressante à approfondir dans le futur serait la création de liaison (communication) entre une méthode d'indexation classique et une méthode basée sur l'apprentissage profond comme illustré dans la figure 7.1, notamment en injectant des descripteurs sous la forme simple de sacs-de-mots à différents niveaux d'un système d'apprentissage profond ou inversement en injectant des descripteurs obtenus par apprentissage profond dans un système d'apprentissage classique. Cette dernière idée a été partiellement intégré dans le système d'indexation proposé par le groupe IRIM soumis à TRECVid 2014 et les premiers résultats obtenus ont été encourageants. Enfin, une fusion de ces deux processus bien distincts est également envisageable.

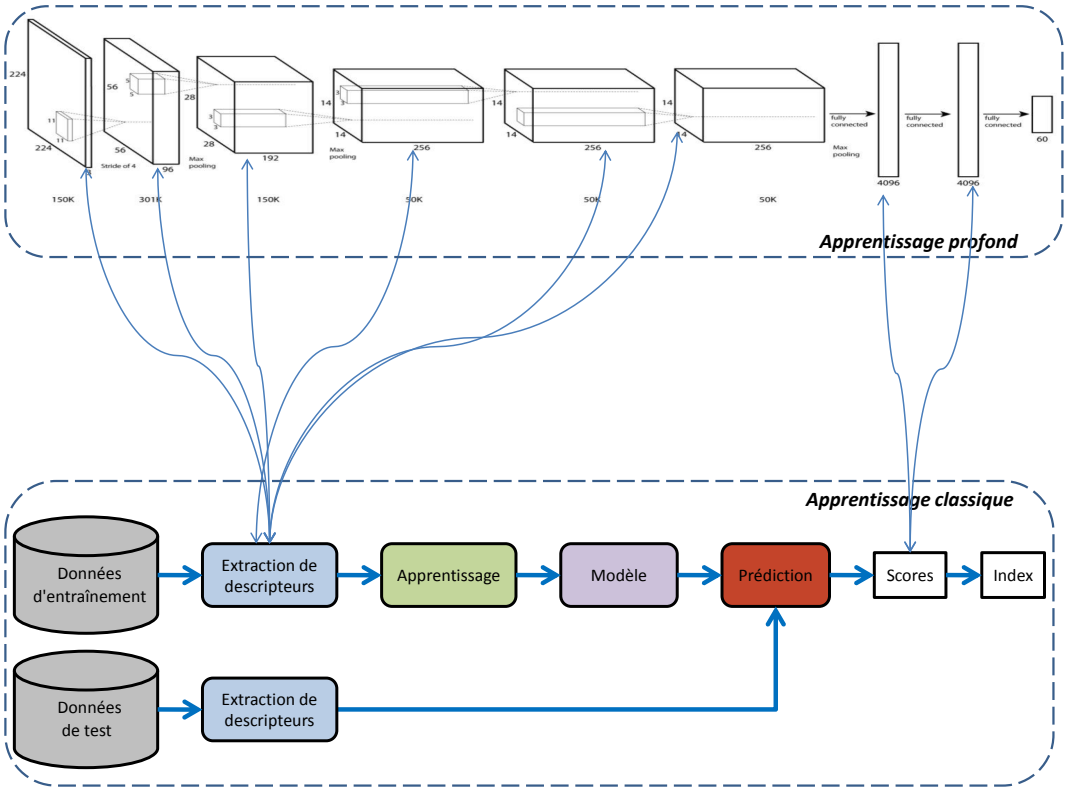


FIGURE 7.1 – La création de liaison entre une méthode d'indexation classique et une méthode basée sur l'apprentissage profond.

Liste des publications

Journal scientifique

2014 Bahjat Safadi, Nadia Derbas, Georges Quénot. *Descriptor Optimization for Multimedia Indexing and Retrieval.* Multimedia Tools and Applications (MTAP).

Conférences internationales

2014 Nadia Derbas, Georges Quénot. *Joint Audio-Visual Words for Violent Scenes Detection in Movies.* ACM International Conference on Multimedia Retrieval (ICMR).

Conférences nationales

2014 Nadia Derbas, Georges Quénot. *Mots audio-visuels joints pour la détection de scènes violentes dans les vidéos.* Conférence en Recherche d'Information et Applications (CORIA).

2013 Nadia Derbas. *Production d'annotations par plan pour l'indexation des vidéos.* Rencontres Jeunes Chercheurs (RJC).

Workshops internationaux - MediaEval

2013 Nadia Derbas, Bahjat Safadi, Georges Quénot. *LIG at MediaEval 2013 Affect Task : Use of a Generic Method and Joint Audio-Visual Words.* MediaEval 2013 Workshop.

2012 Nadia Derbas, Franck Thollard, Bahjat Safadi, Georges Quénot. *LIG at MediaEval 2012 affect task : use of a generic method.* MediaEval 2012 Workshop.

Workshops internationaux - TRECVid

- *IRIM at TRECVid 2013 : Semantic Indexing and Instance Search.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2013.
- *Quaero at TRECVID 2013 : Semantic Indexing and Instance Search.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2013.
- *IRIM at TRECVid 2012 : Semantic Indexing and Instance Search.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2012.
- *Quaero at TRECVID 2012 : Semantic Indexing.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2012.
- *IRIM at TRECVid 2011 : Semantic Indexing and Instance Search.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2011.
- *Quaero at TRECVID 2011 : Semantic Indexing and Multimedia Event Detection.* TRECVideo Retrieval Evaluation Notebook Papers and Slides 2011.

Bibliographie

[Akat 13] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. "Good Practice in Large-Scale Learning for Image Classification". 2013. 35

[Alah 12] A. Alahi, R. Ortiz, and P. Vandergheynst. "Freak : Fast retina keypoint". In : *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 510–517, Ieee, 2012. 122

[Amit 07] Y. Amit and A. Trouvé. "Pop : Patchwork of parts models for object recognition". *International Journal of Computer Vision*, Vol. 75, No. 2, pp. 267–282, 2007. 66

[Atre 10] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. "Multimodal fusion for multimedia analysis : a survey". *Multimedia systems*, Vol. 16, No. 6, pp. 345–379, 2010. 38

[Ayac 07a] S. Ayache and G. Quénot. "Indexation de documents multimédia par réseaux d'opérateurs.". In : *CORIA*, pp. 385–400, 2007. 29

[Ayac 07b] S. Ayache, G. Quénot, and J. Gensel. "Classifier fusion for SVM-based multimedia semantic indexing". In : *Advances in Information Retrieval*, pp. 494–504, Springer, 2007. 39

[Bai 13] H. Bai, Y. Dong, S. Cen, L. Wang, L. Liu, W. Liu, Y. Bian, C. Huang, N. Zhao, B. Liu, *et al.* "ORANGE LABS BEIJING (FTRDBJ) AT TRECVID 2013 : INSTANCE SEARCH". 2013. 79

[Ball 12] N. Ballas, B. Labbé, A. Shabou, H. Le Borgne, P.-H. Gosselin, M. Redi, B. Merialdo, H. Jégou, J. Delhumeau, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, F. Thollard, N. Derbas, G. Quenot, H. Bredin, M. Cord, B. Gao, C. Zhu, Y. Tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benoit, P. Lambert, T. Strat, J. Razik, S. Paris, H. Glotin, T. N. Trung, D. Petrovska-Delacrétaz, G. Chollet, A. Stoian, and M. Crucianu. "IRIM at TRECVID 2012 : Semantic Indexing and Instance Search". In : *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, 2012. 116

[Ball 13] N. Ballas, B. Labbé, H. Le Borgne, P. Gosselin, M. Redi, B. Merialdo, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, A. Hamadi, B. Safadi, T.-T.-T. Vuong, H. Dong, N. Derbas, G. Quénot, B. Gao, C. Zhu,

Y. Tang, E. Dellandrea, C.-E. Bichot, L. Chen, A. Benoît, P. Lambert, and T. Strat. "IRIM at TRECVID 2013 : Semantic Indexing and Instance Search". In : *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, 2013. 74

[Batr 08] D. Batra, T. Chen, and R. Sukthankar. "Space-time shapelets for action recognition". In : *Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*, pp. 1–6, IEEE, 2008. 26

[Bay 08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. "Speeded-up robust features (SURF)". *Computer vision and image understanding*, Vol. 110, No. 3, pp. 346–359, 2008. 22

[Beal 03] M. J. Beal, N. Jojic, and H. Attias. "A graphical model for audiovisual object tracking". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 25, No. 7, pp. 828–836, 2003. 50

[Beng 09] Y. Bengio. "Learning deep architectures for AI". *Foundations and trends® in Machine Learning*, Vol. 2, No. 1, pp. 1–127, 2009. 40

[Berm 11] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar. "Violence Detection in Video Using Computer Vision Techniques". In : *Computer Analysis of Images and Patterns*, pp. 332–339, Springer Berlin Heidelberg, 2011. 51

[Bern 05] E. Bernstein and Y. Amit. "Part-based statistical models for object classification and detection". In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 734–740 vol. 2, 2005. 66

[Bied 87] I. Biederman. "Recognition-by-components : a theory of human image understanding.". *Psychological review*, Vol. 94, No. 2, p. 115, 1987. 19, 20

[Bish 06] C. M. Bishop *et al.* *Pattern recognition and machine learning*. Vol. 1, springer New York, 2006. 104

[Blan 05] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. "Actions as space-time shapes". In : *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 1395–1402, IEEE, 2005. 26

[Bobi 01] A. F. Bobick and J. W. Davis. "The recognition of human movement using temporal templates". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 23, No. 3, pp. 257–267, 2001. 25

[Bore 96] J. S. Boreczky and L. A. Rowe. "Comparison of video shot boundary detection techniques". *Journal of Electronic Imaging*, Vol. 5, No. 2, pp. 122–128, 1996. 13

[Bosc 07] A. Bosch, A. Zisserman, and X. Munoz. "Representing shape with a spatial pyramid kernel". In : *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 401–408, ACM, 2007. 66

[Bosc 08] A. Bosch, A. Zisserman, and X. Muoz. "Scene classification using a hybrid generative/discriminative approach". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 30, No. 4, pp. 712–727, 2008. 22

[Bott 07] L. Bottou and O. Bousquet. "The Tradeoffs of Large Scale Learning.". In : *NIPS*, p. 2, 2007. 36

[Chan 01] C. Chang and C. Lin. "LIBSVM : a library for support vector machines, Software". 2001. 31

[Chan 11] C.-C. Chang and C.-J. Lin. "LIBSVM : a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 2, No. 3, p. 27, 2011. 36

[Chaw 11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE : synthetic minority over-sampling technique". *arXiv preprint arXiv :1106.1813*, 2011. 36

[Chop 05] S. Chopra, R. Hadsell, and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification". In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 539–546, IEEE, 2005. 34

[Chua 02] T.-S. Chua, L. Chen, and J. Wang. "Stratification approach to modeling video". *Multimedia Tools and Applications*, Vol. 16, No. 1-2, pp. 79–97, 2002. 12

[Chum 07] O. Chum and A. Zisserman. "An exemplar model for learning object classes". In : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007. 66

[Como 94] P. Comon. "Independent component analysis, a new concept ?". *Signal processing*, Vol. 36, No. 3, pp. 287–314, 1994. 32

[Cort 95] C. Cortes and V. Vapnik. "Support-vector networks". *Machine learning*, Vol. 20, No. 3, pp. 273–297, 1995. 35

[Cove 67] T. Cover and P. Hart. "Nearest neighbor pattern classification". *Information Theory, IEEE Transactions on*, Vol. 13, No. 1, pp. 21–27, 1967. 34

[Cram 02] K. Crammer and Y. Singer. "On the algorithmic implementation of multiclass kernel-based vector machines". *The Journal of Machine Learning Research*, Vol. 2, pp. 265–292, 2002. 35

[Cran 06] D. J. Crandall and D. P. Huttenlocher. "Weakly supervised learning of part-based spatial models for visual object recognition". In : *Computer Vision–ECCV 2006*, pp. 16–29, Springer, 2006. 65

[Cris 07] M. Cristani, M. Bicego, and V. Murino. "Audio-Visual Event Recognition in Surveillance Video Sequences". *Multimedia, IEEE Transactions on*, Vol. 9, No. 2, pp. 257–267, 2007. 50

[Csur 04] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. "Visual categorization with bags of keypoints". In : *Workshop on statistical learning in computer vision, ECCV*, pp. 1–2, 2004. 23, 102

[Dala 05] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 886–893, IEEE, 2005. 18, 19

[Dala 06] N. Dalal, B. Triggs, and C. Schmid. "Human detection using oriented histograms of flow and appearance". In : *Computer Vision–ECCV 2006*, pp. 428–441, Springer, 2006. 27

[Dana 07] S. Danafar and N. Gheissari. "Action recognition for surveillance applications using optic flow and SVM". In : *Computer Vision–ACCV 2007*, pp. 457–466, Springer, 2007. 27

[Datt 02] A. Datta, M. Shah, and N. da Vitoria Lobo. "Person-on-person violence detection in video data". In : *Pattern Recognition*, pp. 433–438 vol.1, 2002. 51

[Dave 91] G. Davenport, T. A. Smith, and N. Pincever. "Cinematic primitives for multimedia". Vol. 11(4), pp. 67–74, 1991. 12

[Dehg 13] A. Dehghan, G. Shu, N. Souli, W. Li, S. Pehlivan, M. Shah, J. Liu, and H. Cheng. "UCF-CRCV at TRECVID 2013 : Semantic Indexing". 2013. 79

[Dele 11] B. Delezoide, F. Precioso, P.-H. Gosselin, M. Redi, B. Merialdo, L. Granjon, D. Pellerin, M. Rombaut, H. Jégou, R. Vieux, B. Mansencal, J. Benois-Pineau, S. Ayache, B. Safadi, F. Thollard, G. Quénot, H. Bredin, M. Cord, A. Benoit, P. Lambert, T. Strat, J. Razik, S. Paris, and H. Glotin. "IRIM at TRECVID 2011 : Semantic Indexing and Instance Search". In : *Proceedings of the workshop on TREC Video Retrieval Evaluation (TRECVID)*, 2011. 87, 89, 90

[Delh 13] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez. "Revisiting the VLAD image representation". In : *Proceedings of the 21st ACM international conference on Multimedia*, pp. 653–656, ACM, 2013. 25

[Dema 13] C.-H. Demarty, C. Penet, M. Schedl, B. Ionescu, V. L. Quang, and Y.-G. Jiang. "The MediaEval 2013 Affect Task : Violent Scenes Detection". In : *MediaEval Workshop*, Barcelona, Spain, October 18-19 2013. 45, 51, 54

[Dema 14] H. Demarty, B. Ionescu, Y.-G. Jiang, V. L. Quang, M. Schedl, and C. Penet. "Benchmarking Violent Scenes Detection in Movies". In : *IEEE International Workshop on Content-Based Multimedia Indexing-CBMI 2014, 18-20 June, Klagenfurt, Austria*, 2014. 59

[Derb 12] N. Derbas, F. Thollard, B. Safadi, and G. Quénot. "LIG at MediaEval 2012 Affect Task : Use of a Generic Method". In : *MediaEval Workshop*, Pisa, Italy, October 4-5 2012. 38

[Derb 13] N. Derbas, B. Safadi, and G. Quénot. "LIG at MediaEval 2013 Affect Task : Use of a Generic Method and Joint Audio-Visual Words". In : *MediaEval Workshop*, Barcelona, Spain, October 18-19 2013. 58

[Diet 97] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". *Artificial Intelligence*, Vol. 89, No. 1, pp. 31–71, 1997. 66

[Doll 05] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. "Behavior recognition via sparse spatio-temporal features". In : *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pp. 65–72, IEEE, 2005. 27

[Domi 99] P. Domingos. "Metacost : A general method for making classifiers cost-sensitive". In : *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164, ACM, 1999. 36

[Drum 03] C. Drummond, R. C. Holte, *et al.* "C4. 5, class imbalance, and cost sensitivity : why under-sampling beats over-sampling". In : *Workshop on Learning from Imbalanced Datasets II*, Citeseer, 2003. 36

[Dumo 12] E. Dumont and G. Quénot. "A local temporal context-based approach for TV news story segmentation". In : *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 973–978, IEEE, 2012. 13

[Efro 03] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. "Recognizing action at a distance". In : *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 726–733, IEEE, 2003. 27

[Elka 01] C. Elkan. "The foundations of cost-sensitive learning". In : *International joint conference on artificial intelligence*, pp. 973–978, Citeseer, 2001. 36

[Ever 10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. "The pascal visual object classes (voc) challenge". *International journal of computer vision*, Vol. 88, No. 2, pp. 303–338, 2010. 42

[Ever 11] M. Everingham and J. Winn. "The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Development Kit". 2011. 42

[Fara 13] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. "Learning hierarchical features for scene labeling". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 35, No. 8, pp. 1915–1929, 2013. 40

[Farn 03] G. Farnebäck. "Two-frame motion estimation based on polynomial expansion". In : *Image Analysis*, pp. 363–370, Springer, 2003. 21

[Fei 07] L. Fei-Fei, R. Fergus, and A. Torralba. "Recognizing and learning object categories". *CVPR Short Course*, Vol. 2, 2007. 23

[Felz 10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. "Object detection with discriminatively trained part-based models". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 32, No. 9, pp. 1627–1645, 2010. 66

[Feng 12] B. Feng, P. Ding, J. Chen, J. Bai, S. Xu, and B. Xu. "Multi-modal information fusion for news story segmentation in broadcast video". In : *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 1417–1420, IEEE, 2012. 13

[Ferg 07] R. Fergus, P. Perona, and A. Zisserman. "Weakly supervised scale-invariant learning of models for visual recognition". *International Journal of Computer Vision*, Vol. 71, No. 3, pp. 273–303, 2007. 65

[Ferr 08] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. "Groups of adjacent contour segments for object detection". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 30, No. 1, pp. 36–51, 2008. 66

[Fish 00] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. A. Viola. "Learning joint statistical models for audio-visual fusion and segregation". In : *NIPS*, pp. 772–778, 2000. 32

[Fran 08] V. Franc and S. Sonnenburg. "Optimized cutting plane algorithm for support vector machines". In : *Proceedings of the 25th international conference on Machine learning*, pp. 320–327, ACM, 2008. 36

[Fuss 06] M. Fussenegger, A. Opelt, and A. Pinz. "Object localization/segmentation using generic shape priors". In : *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, pp. 41–44, IEEE, 2006. 65

[Gall 08] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. "Weakly supervised object localization with stable segmentations". In : *Computer Vision–ECCV 2008*, pp. 193–207, Springer, 2008. 65

[Geme 08] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. "Kernel codebooks for scene categorization". In : *Computer Vision–ECCV 2008*, pp. 696–709, Springer, 2008. 22, 25

[Geng 05] X. Geng, D.-C. Zhan, and Z.-H. Zhou. "Supervised nonlinear dimensionality reduction for visualization and classification". *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, Vol. 35, No. 6, pp. 1098–1107, 2005. 33

[Geve 99] T. Gevers and A. W. Smeulders. "Color-based object recognition". *Pattern recognition*, Vol. 32, No. 3, pp. 453–464, 1999. 17

[Gian 06] T. Giannakopoulos, D. I. Kosmopoulos, A. Aristidou, and S. Theodoridis. "Violence Content Classification Using Audio Features". In : *SETN*, pp. 502–507, 2006. 51

[Gian 10] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis. "Audio-Visual Fusion for Detecting Violent Scenes in Videos". In : *Artificial Intelligence : Theories, Models and Applications*, pp. 91–100, Springer Berlin Heidelberg, 2010. 38

[Gong 08] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. "Detecting Violent Scenes in Movies by Auditory and Visual Cues". In : *Advances in Multimedia Information Processing - PCM 2008*, pp. 317–326, Springer Berlin Heidelberg, 2008. 38, 51

[Gore 07] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. "Actions as space-time shapes". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 29, No. 12, pp. 2247–2253, 2007. 26

[Gori 10] D. Gorisse, F. Precioso, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, H. Lachambre, E. El Khoury, *et al.* "IRIM at TRECVID 2010 : High level feature extraction and instance search". In : *TREC Video Retrieval Evaluation workshop*, 2010. 107

[Gran 08] D. Grangier and S. Bengio. "A discriminative kernel-based approach to rank images from text queries". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 30, No. 8, pp. 1371–1384, 2008. 35

[Gu 08] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. "Multi-layer multi-instance learning for video concept detection". *Multimedia, IEEE Transactions on*, Vol. 10, No. 8, pp. 1605–1616, 2008. 83

[Habi 14a] A. Habibian, T. Mensink, and C. G. Snoek. "Composite Concept Discovery for Zero-Shot Video Event Detection". 2014. 30

[Habi 14b] A. Habibian and C. G. Snoek. "Stop-Frame Removal Improves Web Video Classification". In : *Proceedings of International Conference on Multimedia Retrieval*, p. 499, ACM, 2014. 83

[Hama 12] A. Hamadi, G. Quénot, and P. Mulhem. "Two-layers re-ranking approach based on contextual information for visual concepts detection in videos". In : *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, pp. 1–6, IEEE, 2012. 114

[Hama 13] A. Hamadi, B. Safadi, T.-T.-T. Vuong, D. Han, N. Derbas, P. Mulhem, and G. Quénot. "Quaero at TRECVID 2013 : Semantic Indexing and Instance Search". In : *Proc. TRECVID Workshop*, Gaithersburg, MD, USA, nov 2013. 74

[Hans 00] L. K. Hansen, J. Larsen, and T. Kolenda. "On independent component analysis for multimedia signals". *Multimedia Image and Video Processing*, pp. 175–199, 2000. 32

[Harr 54] Z. S. Harris. "Distributional structure.". *Word*, 1954. 23

[Harr 88] C. Harris and M. Stephens. "A combined corner and edge detector.". In : *Alvey vision conference*, p. 50, Manchester, UK, 1988. 19

[Harz 09] H. Harzallah, F. Jurie, and C. Schmid. "Combining efficient object localization and image classification". In : *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 237–244, IEEE, 2009. 66

[Huan 97] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. "Image indexing using color correlograms". In : *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 762–768, IEEE, 1997. 18

[Jego 10a] H. Jégou, M. Douze, and C. Schmid. "Improving bag-of-features for large scale image search". *International Journal of Computer Vision*, Vol. 87, No. 3, pp. 316–336, 2010. 25

[Jego 10b] H. Jégou, M. Douze, C. Schmid, and P. Pérez. "Aggregating local descriptors into a compact image representation". In : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3304–3311, IEEE, 2010. 25

[Jego 12a] H. Jégou and O. Chum. "Negative evidences and co-occurences in image retrieval : The benefit of PCA and whitening". In : *Computer Vision–ECCV 2012*, pp. 774–787, Springer, 2012. 104

[Jego 12b] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. "Aggregating local image descriptors into compact codes". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 34, No. 9, pp. 1704–1716, 2012. 25, 32

[Jia 13] Y. Jia. "Caffe : An Open Source Convolutional Architecture for Fast Feature Embedding". http://caffe.berkeleyvision.org/, 2013. 40

[Jian 09] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. "Domain adaptive semantic diffusion for large scale context-based video annotation". In : *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1420–1427, IEEE, 2009. 29

[Jian 11] W. Jiang and A. C. Loui. "Audio-visual grouplet : temporal audio-visual interactions for general video concept classification". In : *ACM Multimedia*, pp. 123–132, 2011. 51

[Jian 14] L. Jiang, W. Tong, D. Meng, and A. G. Hauptmann. "Towards Efficient Learning of Optimal Spatial Bag-of-Words Representations". 2014. 25

[Joac 99] T. Joachims. "Making large scale SVM learning practical". 1999. 36

[Joll 05] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. 32, 33

[Kank 00] M. S. Kankanhalli and T.-S. Chua. "Video modeling using strata-based annotation". *Multimedia, IEEE*, Vol. 7, No. 1, pp. 68–74, 2000. 12

[Kapl 97] L. M. Kaplan, R. Murenzi, and K. R. Namuduri. "Fast texture database retrieval using extended fractal features". In : *Photonics West'98 Electronic Imaging*, pp. 162–173, International Society for Optics and Photonics, 1997. 18

[Kell 08] V. Kellokumpu, G. Zhao, and M. Pietikäinen. "Human activity recognition using a dynamic texture based method.". In : *BMVC*, pp. 1–10, 2008. 27

[Khou 14] E. el Khoury, C. Sénac, and P. Joly. "Audiovisual diarization of people in video content". *Multimedia Tools Appl.*, Vol. 68, No. 3, pp. 747–775, 2014. 13

[Kole 02] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther. "Independent component analysis for understanding multimedia content". In : *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pp. 757–766, IEEE, 2002. 32

[Kriz 12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In : *Advances in neural information processing systems*, pp. 1097–1105, 2012. 40, 41, 122

[Kuba 97] M. Kubat, S. Matwin, *et al.* "Addressing the curse of imbalanced training sets : one-sided selection". In : *ICML*, pp. 179–186, 1997. 36

[Kueh 11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. "HMDB : a large video database for human motion recognition". In : *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 2556–2563, IEEE, 2011. 43

[La C 98] M. La Cascia, S. Sethi, and S. Sclaroff. "Combining textual and visual cues for content-based image retrieval on the world wide web". In : *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*, pp. 24–28, IEEE, 1998. 32

[Lamp 08] C. H. Lampert, M. B. Blaschko, and T. Hofmann. "Beyond sliding windows : Object localization by efficient subwindow search". In : *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008. 66

[Lan 13] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. Hauptmann. "Multimedia classification and event detection using double fusion". *Multimedia Tools and Applications*, pp. 1–15, 2013. 39

[Lapt 05] I. Laptev. "On space-time interest points". *International Journal of Computer Vision*, Vol. 64, No. 2-3, pp. 107–123, 2005. 27, 28, 55, 56, 89

[Laze 05] S. Lazebnik, C. Schmid, and J. Ponce. "A sparse texture representation using local affine regions". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 27, No. 8, pp. 1265–1278, 2005. 22

[Laze 06] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories". In : *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 2169–2178, IEEE, 2006. 25

[LeCu 10] Y. LeCun, K. Kavukcuoglu, and C. Farabet. "Convolutional networks and applications in vision". In : *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pp. 253–256, IEEE, 2010. 40

[LeCu 98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998. 36

[Lee 04] Y. Lee, Y. Lin, and G. Wahba. "Multicategory support vector machines : Theory and application to the classification of microarray data and satellite radiance data". *Journal of the American Statistical Association*, Vol. 99, No. 465, pp. 67–81, 2004. 35

[Leib 04] B. Leibe, A. Leonardis, and B. Schiele. "Combined object categorization and segmentation with an implicit shape model". In : *Workshop on Statistical Learning in Computer Vision, ECCV*, p. 7, 2004. 66

[Li 02] F. F. Li, R. VanRullen, C. Koch, and P. Perona. "Rapid natural scene categorization in the near absence of attention". *Proceedings of the National Academy of Sciences*, Vol. 99, No. 14, pp. 9596–9601, 2002. 15

[Li 10] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li. "Object Bank : A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification.". In : *NIPS*, p. 5, 2010. 29, 30

[Lin 09] J. Lin and W. Wang. "Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training". In : *Advances in Multimedia Information Processing - PCM 2009*, pp. 930–935, Springer Berlin Heidelberg, 2009. 38

[Lin 11] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. "Large-scale image classification : fast feature extraction and svm training". In : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1689–1696, IEEE, 2011. 25, 36

[Lind 98] T. Lindeberg. "Feature detection with automatic scale selection". *International journal of computer vision*, Vol. 30, No. 2, pp. 79–116, 1998. 19

[Liu 08] D. Liu, G. Hua, P. Viola, and T. Chen. "Integrated feature selection and higher-order spatial feature extraction for object categorization". In : *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008. 25

[Liu 09] X.-Y. Liu, J. Wu, and Z.-H. Zhou. "Exploratory undersampling for class-imbalance learning". *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, Vol. 39, No. 2, pp. 539–550, 2009. 36

[Lowe 04] D. G. Lowe. "Distinctive image features from scale-invariant keypoints". *International journal of computer vision*, Vol. 60, No. 2, pp. 91–110, 2004. 20, 21, 22, 89

[Lowe 99] D. G. Lowe. "Object recognition from local scale-invariant features". In : *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, pp. 1150–1157, Ieee, 1999. 21

[Lu 10] M.-M. Lu, L. Xie, Z.-H. Fu, D.-M. Jiang, and Y.-N. Zhang. "Multimodal feature integration for story boundary detection in broadcast news". In : *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, pp. 420–425, IEEE, 2010. 13

[Luca 81] B. D. Lucas, T. Kanade, *et al.* "An iterative image registration technique with an application to stereo vision.". In : *IJCAI*, pp. 674–679, 1981. 21

[Manj 96] B. S. Manjunath and W.-Y. Ma. "Texture features for browsing and retrieval of image data". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 18, No. 8, pp. 837–842, 1996. 18

[Mao 92] J. Mao and A. K. Jain. "Texture classification and segmentation using multiresolution simultaneous autoregressive models". *Pattern recognition*, Vol. 25, No. 2, pp. 173–188, 1992. 18

[Mare 05] R. Maree, P. Geurts, J. Piater, and L. Wehenkel. "Random subwindows for robust image classification". In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pp. 34–40, IEEE, 2005. 20

[Mars 09] M. Marszalek, I. Laptev, and C. Schmid. "Actions in context". In : *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE, 2009. 43

[Mazl 13] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. "Searching informative concept banks for video event detection". In : *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp. 255–262, ACM, 2013. 30

[Meek 13] M. Meeker and L. Wu. "2013 internet trends". *Kleiner Perkins Caufield & Byers, Technical Report*, 2013. 2

[Merl 12] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. "Semantic model vectors for complex video event recognition". *Multimedia, IEEE Transactions on*, Vol. 14, No. 1, pp. 88–101, 2012. 30

[Miko 04] K. Mikolajczyk and C. Schmid. "Scale & affine invariant interest point detectors". *International journal of computer vision*, Vol. 60, No. 1, pp. 63–86, 2004. 19

[Miko 05] K. Mikolajczyk and C. Schmid. "A performance evaluation of local descriptors". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 27, No. 10, pp. 1615–1630, 2005. 21

[Mitr 10] D. Mitrović, M. Zeppelzauer, and C. Breiteneder. "Features for content-based audio retrieval". *Advances in computers*, Vol. 78, pp. 71–150, 2010. 28

[Muhl 12] M. Mühling, R. Ewerth, J. Zhou, and B. Freisleben. "Multimodal video concept detection via bag of auditory words and multiple kernel learning". In : *Advances in Multimedia Modeling*, pp. 40–50, Springer, 2012. 39

[Naph 04] M. R. Naphade and J. R. Smith. "On the detection of semantic concepts at TRECVID". In : *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 660–667, ACM, 2004. 12

[Nguy 09] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. "Weakly supervised discriminative localization and classification : a joint learning process". In : *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1925–1932, IEEE, 2009. 65

[Nowa 06] E. Nowak, F. Jurie, and B. Triggs. "Sampling strategies for bag-of-features image classification". In : *Computer Vision–ECCV 2006*, pp. 490–503, Springer, 2006. 20

[Ojal 02] T. Ojala, M. Pietikainen, and T. Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 24, No. 7, pp. 971–987, 2002. 18

[Ojal 96] T. Ojala, M. Pietikäinen, and D. Harwood. "A comparative study of texture measures with classification based on featured distributions". *Pattern recognition*, Vol. 29, No. 1, pp. 51–59, 1996. 18

[Oliv 01] A. Oliva and A. Torralba. "Modeling the shape of the scene : A holistic representation of the spatial envelope". *International journal of computer vision*, Vol. 42, No. 3, pp. 145–175, 2001. 18

[Opel 05] A. Opelt and A. Pinz. "Object localization with boosting and weak supervision for generic object recognition". In : *Image Analysis*, pp. 862–871, Springer, 2005. 65

[Over 05] P. Over, T. Ianeva, W. Kraaij, and A. F. Smeaton. "TRECVID 2005-an overview". 2005. 12

[Over 13] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, *et al.* "An overview of the goals, tasks, data, evaluation mechanisms and metrics". In : *TRECVID 2013-TREC Video Retrieval Evaluation Online*, 2013. 45, 73, 74

[Pand 11] M. Pandey and S. Lazebnik. "Scene recognition and weakly supervised object localization with deformable part-based models". In : *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1307–1314, IEEE, 2011. 65

[Pari 10] D. Parikh and C. L. Zitnick. "The role of features, algorithms and data in visual recognition". In : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2328–2335, IEEE, 2010. 16

[Pear 01] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Vol. 2, No. 11, pp. 559–572, 1901. 32

[Pene 13] C. Penet, C.-H. Demarty, G. Gravier, and P. Gros. "Audio event detection in movies using multiple audio words and contextual Bayesian networks". In : *Workshop on Content-Based Multimedia Indexing*, pp. 17–22, 2013. 51

[Perr 07] F. Perronnin and C. Dance. "Fisher kernels on visual vocabularies for image categorization". In : *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007. 25

[Perr 10a] F. Perronnin, J. Sánchez, and Y. Liu. "Large-scale image categorization with explicit data embedding". In : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2297–2304, IEEE, 2010. 36

[Perr 10b] F. Perronnin, J. Sánchez, and T. Mensink. "Improving the fisher kernel for large-scale image classification". In : *Computer Vision–ECCV 2010*, pp. 143–156, Springer, 2010. 25, 31, 32, 102

[Plat 99] J. C. Platt. "Fast training of support vector machines using sequential minimal optimization". 1999. 36

[Pres 12] A. Prest, C. Schmid, and V. Ferrari. "Weakly supervised learning of interactions between humans and objects". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 34, No. 3, pp. 601–614, 2012. 65

[Quac 07] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. "Efficient mining of frequent and distinctive feature configurations". In : *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007. 25

[Quen 01] G. Quénot. "TREC-10 Shot Boundary Detection Task : CLIPS System Description and Evaluation.". In : *TREC*, 2001. 13

[Quen 12] G. Quénot and F. Thollard. "Reclassement d'images par le contenu". In : *CORIA 2012*, Bordeaux, mar 2012. 83, 85, 96

[Rama 06] D. Ramanan and C. Sminchisescu. "Training Deformable Models for Localization". In : *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 206–213, 2006. 66

[Redi 11] M. Redi and B. Merialdo. "Saliency moments for image categorization". In : *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, p. 39, ACM, 2011. 107

[Ries 12] C. X. Ries and R. Lienhart. "Deriving a discriminative color model for a given object class from weakly labeled training data". In : *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, p. 44, ACM, 2012. 66, 68

[Rohr 11] M. Rohrbach, M. Stark, and B. Schiele. "Evaluating knowledge transfer and zero-shot learning in a large-scale setting". In : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1641–1648, IEEE, 2011. 36

[Rowe 00] S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding". *Science*, Vol. 290, No. 5500, pp. 2323–2326, 2000. 33

[Rowl 95] H. A. Rowley, S. Baluja, T. Kanade, *et al. Human face detection in visual scenes*. School of Computer Science, Carnegie Mellon University, 1995. 66

[Rui 98] Y. Rui, T. S. Huang, and S. Mehrotra. "Exploring Video Structure Beyond The Shots". In : *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, pp. 237–, IEEE Computer Society, Washington, DC, USA, 1998. 12

[Russ 06] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. "Using multiple segmentations to discover objects and their extent in image collections". In : *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 1605–1614, IEEE, 2006. 65

[Sada 12] S. Sadanand and J. J. Corso. "Action bank : A high-level representation of activity in video". In : *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1234–1241, IEEE, 2012. 30

[Safa 10] B. Safadi and G. Quénot. "Evaluations of multi-learner approaches for concept indexing in video documents". In : *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pp. 88–91, 2010. 36, 37, 55, 103, 108

[Safa 11a] B. Safadi, N. Derbas, A. Hamadi, F. Thollard, G. Quénot, H. Jégou, T. Gehrig, H. K. Ekenel, R. Stifelhagen, *et al.* "Quaero at TRECVID 2011 : Semantic Indexing and Multimedia Event Detection". In : *TRECVID 2011-TREC Video Retrieval Evaluation workshop*, 2011. 32

[Safa 11b] B. Safadi and G. Quenot. "Re-ranking for multimedia indexing and retrieval". In : *Advances in Information Retrieval*, pp. 708–711, Springer, 2011. 15, 75, 114

[Safa 12] B. Safadi and G. Quénot. "Active learning with multiple classifiers for multimedia indexing". *Multimedia Tools and Applications*, Vol. 60, No. 2, pp. 403–417, 2012. 37

[Sanc 11] J. Sánchez and F. Perronnin. "High-dimensional signature compression for large-scale image classification". In : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1665–1672, IEEE, 2011. 36

[Sanc 13] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. "Image classification with the Fisher vector : Theory and practice". *International journal of computer vision*, Vol. 105, No. 3, pp. 222–245, 2013. 103

[Sarg 06] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp. "Multimodal speaker identification using canonical correlation analysis". In : *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, pp. I–I, IEEE, 2006. 51

[Sarg 07] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp. "Audiovisual synchronization and fusion using canonical correlation analysis". *Multimedia, IEEE Transactions on*, Vol. 9, No. 7, pp. 1396–1403, 2007. 51

[Schu 04] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions : a local SVM approach". In : *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, pp. 32–36, IEEE, 2004. 41

[Shal 08] S. Shalev-Shwartz and N. Srebro. "SVM optimization : inverse dependence on training set size". In : *Proceedings of the 25th international conference on Machine learning*, pp. 928–935, ACM, 2008. 36

[Sivi 03] J. Sivic and A. Zisserman. "Video Google : A text retrieval approach to object matching in videos". In : *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1470–1477, IEEE, 2003. 23, 102

[Smar 03] P. Smaragdis and M. Casey. "Audio/visual independent components". In : *Proc. ICA*, pp. 709–714, 2003. 32

[Smea 10] A. F. Smeaton, P. Over, and A. R. Doherty. "Video shot boundary detection : Seven years of TRECVid activity". *Computer Vision and Image Understanding*, Vol. 114, No. 4, pp. 411–418, 2010. 13

[Smit 03] J. R. Smith, M. Naphade, and A. Natsev. "Multimedia semantic indexing using model vectors". In : *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, pp. II–445, IEEE, 2003. 29

[Smit 93] T. Smith and G. Davenport. “The stratification system a design environment for random access video”. In : P. Venkat Rangan, Ed., *Network and Operating System Support for Digital Audio and Video*, pp. 250–261, Springer Berlin Heidelberg, 1993. 12

[Snoe 05] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. “Early Versus Late Fusion in Semantic Video Analysis”. In : *ACM International Conference on Multimedia*, pp. 399—-402, 2005. 38

[Snoe 13] C. Snoek, K. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S. Pintea, R. Tao, *et al.* “MediaMill at TRECVID 2013 : Searching concepts, objects, instances and events in video”. In : *NIST TRECVID Workshop*, 2013. 41, 79

[Souz 10] F. de Souza, G. Chávez, E. do Valle, and A. de A Araujo. “Violence Detection in Video Using Spatio-Temporal Features”. In : *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 224–230, Washington, DC, USA, August 30-Septembre 3 2010. 51

[Stra 14] S. T. Strat, A. Benoit, P. Lambert, and A. Caplier. “Retina enhanced SURF descriptors for spatio-temporal concept detection”. *Multimedia tools and applications*, Vol. 69, No. 2, pp. 443–469, 2014. 122

[Stri 95] M. A. Stricker and M. Orengo. “Similarity of color images”. In : *IS&T/SPIE’s Symposium on Electronic Imaging : Science & Technology*, pp. 381–392, International Society for Optics and Photonics, 1995. 18

[Swai 91] M. J. Swain and D. H. Ballard. “Color indexing”. *International journal of computer vision*, Vol. 7, No. 1, pp. 11–32, 1991. 17

[Tahi 09] M. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. E. van de Sande, and T. Gevers. “Visual category recognition using spectral regression and kernel discriminant analysis”. In : *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pp. 178–185, IEEE, 2009. 36

[Thur 08] C. Thurau and V. Hlavác. “Pose primitive based human action recognition in videos or still images”. In : *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008. 27

[Ting 00] K. M. Ting. *An empirical study of metacost using boosting algorithms*. Springer, 2000. 36

[Todo 06] S. Todorovic and N. Ahuja. “Extracting subimages of an unknown category from a set of images”. In : *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pp. 927–934, IEEE, 2006. 65

[Torr 10] L. Torresani, M. Szummer, and A. Fitzgibbon. "Efficient object category recognition using classemes". In : *Computer Vision–ECCV 2010*, pp. 776–789, Springer, 2010. 29

[Turn 86] M. R. Turner. "Texture discrimination by Gabor functions". *Biological Cybernetics*, Vol. 55, No. 2-3, pp. 71–82, 1986. 18

[Tuyt 10] T. Tuytelaars. "Dense interest points". In : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2281–2288, IEEE, 2010. 20, 21

[Ulge 08] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. "Identifying relevant frames in weakly labeled videos for training concept detectors". In : *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pp. 9–16, ACM, 2008. 83

[Van 06] J. Van De Weijer, T. Gevers, and A. D. Bagdanov. "Boosting color saliency in image feature detection". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 28, No. 1, pp. 150–156, 2006. 22

[Van 10] K. E. Van De Sande, T. Gevers, and C. G. Snoek. "Evaluating color descriptors for object and scene recognition". *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 32, No. 9, pp. 1582–1596, 2010. 25, 59, 69, 108

[Vino 03] A. Vinokourov, D. R. Hardoon, and J. Shawe-Taylor. "Learning the semantics of multimedia content with application to web image retrieval and classification". 2003. 32

[Wan 98] X. Wan and C.-C. Kuo. "A new approach to image retrieval with hierarchical color clustering". *Circuits and Systems for Video Technology, IEEE Transactions on*, Vol. 8, No. 5, pp. 628–643, 1998. 17

[Wang 11] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. "Action recognition by dense trajectories". In : *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3169–3176, IEEE, 2011. 27, 40

[Wein 06] D. Weinland, R. Ronfard, and E. Boyer. "Free viewpoint action recognition using motion history volumes". *Computer Vision and Image Understanding*, Vol. 104, No. 2, pp. 249–257, 2006. 26

[Wein 09] K. Q. Weinberger and L. K. Saul. "Distance metric learning for large margin nearest neighbor classification". *The Journal of Machine Learning Research*, Vol. 10, pp. 207–244, 2009. 34

[Weis 95] R. Weiss, A. Duda, and D. K. Gifford. "Composition and search with a video algebra". *Multimedia, IEEE*, Vol. 2, No. 1, pp. 12–25, 1995. 12

[West 10] J. Weston, S. Bengio, and N. Usunier. "Large scale image annotation : learning to rank with joint word-image embeddings". *Machine learning*, Vol. 81, No. 1, pp. 21–35, 2010. 36

[West 99] J. Weston, C. Watkins, *et al.* "Support vector machines for multi-class pattern recognition.". In : *ESANN*, pp. 61–72, 1999. 35

[Winn 05a] J. Winn, A. Criminisi, and T. Minka. "Object categorization by learned universal visual dictionary". In : *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 1800–1807, IEEE, 2005. 32

[Winn 05b] J. Winn and N. Jojic. "Locus : Learning object classes with unsupervised segmentation". In : *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pp. 756–763, IEEE, 2005. 65

[Yang 10] J. Yang, K. Yu, and T. Huang. "Efficient highly over-complete sparse coding using a mixture model". In : *Computer Vision–ECCV 2010*, pp. 113–126, Springer, 2010. 25

[Ye 12] G. Ye, I.-H. Jhuo, D. Liu, Y.-G. Jiang, D. Lee, and S.-F. Chang. "Joint Audio-Visual Bi-Modal Codewords for Video Event Detection". In : *ACM International Conference on Multimedia Retrieval (ICMR)*, Hong Kong, June 5-8 2012. 50

[Yilm 08] A. Yilmaz and M. Shah. "A differential geometric approach to representing the human actions". *Computer Vision and Image Understanding*, Vol. 109, No. 3, pp. 335–351, 2008. 26

[Zhan 10] Y. Zhang and T. Chen. "Weakly Supervised Object Recognition and Localization with Invariant High Order Features.". In : *BMVC*, pp. 1–11, 2010. 65

[Zhou 06] Z.-H. Zhou and X.-Y. Liu. "Training cost-sensitive neural networks with methods addressing the class imbalance problem". *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 18, No. 1, pp. 63–77, 2006. 36

Abstract

A consequence of the rise of digital technology is that the quantity of available collections of multimedia documents is permanently and strongly increasing. The indexing of these documents became both very costly and impossible to do manually. In order to be able to analyze, classify and search multimedia documents, indexing systems have been defined. However, most of these systems suffer quality or practicability issues. Their performance is limited and depends on the data volume and data variability. Indexing systems analyze multimedia documents, looking for static concepts (bicycle, chair...), or events (wedding, protest...). Therefore, the variability in shapes, positions, lighting or orientation of objects hinders the process. Another aspect is that systems must be scalable. They should be able to handle big data while using reasonable amount of computing time and memory.

The aim of this thesis is to improve the general performance of content-based multimedia indexing systems. Four main contributions are brought in this thesis for improving different stages of the indexing process. The first one is an "early-early fusion method" that merges different information sources in order to extract their deep correlations. This method is used for violent scenes detection in movies. The second contribution is a weakly supervised method for basic concept (objects) localization in images. This can be used afterwards as a new descriptor to help detecting complex concepts (events). The third contribution tackles the noise reduction problem on ambiguously annotated data. Two methods are proposed: a shot annotation generator, and a shot weighing method. The last contribution is a generic descriptor optimization method, based on PCA and non-linear transforms.

These four contributions are tested and evaluated using reference data collections, including TRECVid and MediaEval. These contributions helped our submissions achieving very good rankings in those evaluation campaigns.

Keywords: Multimedia indexing, multimodal fusion, concept localization, annotation filtering, descriptor optimization.

Résumé

L'explosion de la quantité de documents multimédias, suite à l'essor des technologies numériques, a rendu leur l'indexation très coûteuse et manuellement impossible. Par conséquent, le besoin de disposer de systèmes d'indexation capables d'analyser, de stocker et de retrouver les documents multimédias automatiquement, et en se basant sur leur contenu (audio, visuel), s'est fait ressentir dans de nombreux domaines applicatifs. Cependant, les techniques d'indexation actuelles rencontrent encore des problèmes de faisabilité ou de qualité. Leur performance reste très limitée et est dépendante de plusieurs facteurs comme la variabilité et la quantité de données à traiter. En effet, les systèmes d'indexation cherchent à reconnaître des concepts statiques, comme des objets (vélo, chaise, ...), ou des événements (mariage, manifestation, ...). Ces systèmes se heurtent donc au problème de variabilité de formes, de positions, de poses, d'illuminations, d'orientations des objets. Le passage à l'échelle pour pouvoir traiter de très grands volumes de données tout en respectant des contraintes de temps de calcul et de stockage est également une contrainte.

Dans cette thèse, nous nous intéressons à l'amélioration de la performance globale de ces systèmes d'indexation de documents multimédias par le contenu. Pour cela nous abordons le problème sous différents angles et apportons quatre contributions à divers stades du processus d'indexation. Nous proposons tout d'abord une nouvelle méthode de fusion « doublement précoce » entre différentes modalités ou différentes sources d'informations afin d'exploiter au mieux la corrélation entre les modalités. Cette méthode est ensuite appliquée à la détection de scènes violentes dans les films. Nous développons ensuite une méthode faiblement supervisée pour la localisation des concepts basiques (comme les objets) dans les images qui pourra être utilisé plus tard comme un descripteur et une information supplémentaire pour la détection de concepts plus complexes (comme des événements). Nous traitons également la problématique de réduction du bruit généré par des annotations ambiguës sur les données d'apprentissage en proposant deux méthodes : une génération de nouvelles annotations au niveau des plans et une méthode de pondération des plans. Enfin, nous avons mis en place une méthode d'optimisation des représentations du contenu multimédia qui combine une réduction de dimension basée sur une ACP et des transformations non linéaires.

Les quatre contributions sont testées et évaluées sur les collections de données faisant référence dans le domaine, comme TRECVid ou MediaEval. Elles ont participé au bon classement de nos soumissions dans ces campagnes.

Mots Clefs : Indexation multimédia, fusion multimodale, localisation de concepts, filtrage d'annotations, optimisation de descripteurs.