Les données multimédia

Vincent Claveau, Olivier Cappé

▶ **To cite this version:**

Vincent Claveau, Olivier Cappé. Les données multimédia. Les Big Data à découvert, CNRS Éditions, pp.1, 2017, 978-2-271-11464-8. hal-01848660

HAL Id: hal-01848660
https://inria.hal.science/hal-01848660

Submitted on 25 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

6. Les données multimédia

Olivier Cappé et Vincent Claveau

Des signaux à leur interprétation

Les données multimédia (qui mélangent textes, images, sons vidéos...) sont omniprésentes dans notre vie numérique, qu'elle soit professionnelle (vidéosurveillance, MOOC, Web...) ou personnelle (TV, réseaux sociaux, vidéos familiales...). Face à ces quantités de données, les besoins et les opportunités sont multiples : l'utilisateur veut pouvoir stocker les données (quelques jours pour des vidéos de surveillance jusqu'à indéfiniment pour des vidéos personnelles), les indexer pour les retrouver facilement, détecter les duplicatas (fraude aux copyrights, détection de plagiat), les résumer ou les enrichir avec d'autres informations, les présenter sous une autre forme (transcription, traduction, description textuelle d'une vidéo)...L'approche dominante pour aborder ce type d'applications consiste à recourir à des techniques d'apprentissage artificiel pour entraîner un modèle à effectuer la tâche attendue à partir d'exemples de données multimédia, de préférence annotées par un expert (au moins partiellement). Chaque modalité (texte, son, vidéo...) a ses propres particularités, mais les données multimédia sont avant tout destinées à la perception humaine. Toutes les applications évoquées ci-dessus nécessitent donc que l'ordinateur soit capable d'en appréhender le contenu, comme le ferait un humain. Ce passage d'un ensemble de valeurs numériques à une représentation du contenu est l'un des axes principaux de la recherche dans le domaine des données multimédia. Enfin, la qualité du résultat attendu dépend essentiellement du contexte. Ainsi, la génération d'une description textuelle d'une vidéo doit être parfaite si elle est à destination de malvoyants, mais peut supporter quelques erreurs si elle sert de base à l'indexation automatique de la vidéo. L'évaluation dans le domaine du multimédia est donc principalement fondée sur le test sur des données réelles et sur des tâches bien identifiées.

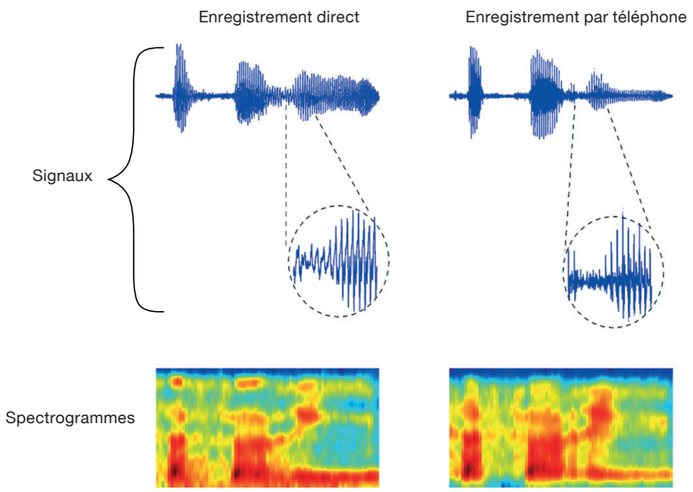

Fig. 1 – **Représentation classique d'un signal audio, dite « spectrogramme », pour un même extrait de parole (le mot « safari ») enregistré dans deux conditions différentes.** Si les formes d'ondes des signaux ne sont pas directement comparables (la zone circulaire correspond à un dixième de seconde de signal), les deux spectrogrammes révèlent clairement des motifs caractéristiques communs. ■

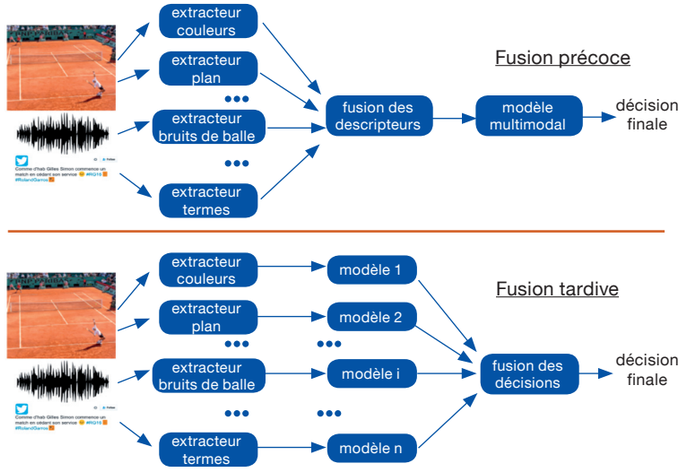

Fig. 2 – **Synchronisation des indices multimodaux d'une vidéo selon deux approches différentes.** La fusion tardive est souvent plus simple à mettre en œuvre que la fusion précoce, mais cette dernière permet de mieux tirer parti de la complémentarité des indices multimodaux. ■

Représentation des données

La caractéristique principale des données brutes correspondant aux signaux multimédia est leur très forte redondance. Les signaux audio (parole, musique) comportent des sections où le signal enregistré apparaît comme périodique (avec un motif qui se répète), les images peuvent contenir de larges zones d'aspect uniforme (aplat de couleur) ou répétitif (texture). Cette redondance est exploitée par les techniques de compression, qui sont particulièrement efficaces pour ce type de données (formats JPEG, MP3...). De surcroît, la perception humaine est robuste vis-à-vis de distorsions ou de dégradations des signaux (bruit de fond ou écho pour les signaux sonores ; flou, pixellisation, distorsion des couleurs pour les images), qui peuvent être relativement élevées sans pour autant remettre en cause le contenu sémantique des données même si leur qualité est altérée de façon perceptible (cf. III.15).

Les représentations (ou descripteurs) typiques des données multimédias doivent donc présenter trois caractéristiques principales. Tout d'abord, elles doivent être compactes, pour éliminer au mieux la redondance mentionnée ci-dessus et assurer un stockage et une manipulation efficaces. Elles doivent également être robustes vis-à-vis de distorsions du signal qui altèrent sa qualité perçue sans pour autant changer son contenu informationnel. Enfin, elles doivent permettre de calculer rapidement des indices de similarité entre signaux, pertinents du point de vue de la perception (figure 1).

Modélisation et multimodalité

Les modèles utilisés pour réaliser les différentes tâches mentionnées en introduction à partir des représentations vues précédemment reposent en général sur des concepts statistiques (cf. IV.8) ainsi que sur l'utilisation de procédures d'optimisation (cf. IV.12). Dans des applications complexes, il est nécessaire de combiner les connaissances que l'on sait extraire de chacun des médias composant les données multimédia. Ainsi, dans une vidéo de tennis, les bruits de balle, les applaudissements et la parole sont extraits de la bande-son ; la vidéo peut être segmentée en plans et chaque plan caractérisé (plan large, gros plan...) ; et les images fournissent d'autres informations (incrustations de scores, présence dominante de vert, reconnaissance de visage ou d'objets...). Des informations externes peuvent aussi être collectées, comme les tweets se rapportant au match. La difficulté est alors de synchroniser ces différents indices dits « multimodaux » au sein d'un modèle pour accéder à une compréhension fine de la vidéo. Pour ce faire, plusieurs approches sont possibles (figure 2) : les représentations de chaque média peuvent être combinées en une représentation globale ensuite utilisée en entrée du modèle (fusion précoce), ou un modèle par média peut être appris et leurs décisions ensuite combinées (fusion tardive).

Plus récemment, l'apprentissage profond (cf. IV.10) a permis d'apprendre directement des représentations multimodales sans reposer sur l'expertise humaine pour définir des représentations adaptées à chaque type de signal. Ce type d'approche ouvre de nombreuses perspectives pour le traitement des données multimédia.

Références bibliographiques

- P. GROS – *L'Indexation multimédia : description et recherche automatiques*, Traité IC2, Hermès-Lavoisier, 2007.
- Conférences TREC – http://trecvid.nist.gov.
- Conférences CLEF – http://www.clef-initiative.eu/web/clef-initiative/home.
- Ateliers MediaEval – http://www.multimediaeval.org.