# Apprentissage pour la synthèse visuelle guidée par un flux audio et application dans le champ de l'art contemporain

Cédric FÉVOTTE[1], Jérôme GRIVEL[2]

[1]Laboratoire Lagrange (CNRS, Observatoire de la Côte d'Azur & Université Nice Sophia Antipolis), Parc Valrose, Nice

[2]Groupe FRAME, Paris

cfevotte@unice.fr, jeromegrivel@hotmail.fr

**Résumé –** Nous présentons une méthode de synthèse d'un flux visuel guidée par un flux audio. Notre approche procède dans une première étape par apprentissage de motifs audiovisuels dans des séquences cinématographiques d'entraînement. L'extraction des motifs est réalisée au moyen d'une co-factorisation non-négative de représentations matricielles des flux audio et visuels. La structure des motifs appris est ensuite utilisée dans une seconde étape pour générer des images à partir de n'importe quelle entrée audio. Ce dispositif a été utilisé dans le cadre d'un projet arts/sciences organisé et exposé par l'Université Nice Sophia Antipolis et l'association pour l'art contemporain DEL'ART.

**Abstract –** We present a method to synthesise a visual stream driven by an audio stream. Our approach relies on the extraction of audiovisual patterns from a collection of short movie excerpts, in a training stage. The extraction is achieved by nonnegative co-factorisation of matrix representations of the audio and visual streams. The structure of the estimated patterns is used in a second stage to synthesise images from any audio input. Our work was carried out in the context of an art/science project organised and exhibited by University Nice Sophia Antipolis and contemporary art organisation DEL'ART.

## 1 Contexte

Le travail rapporté dans cet article s'inscrit dans le cadre d'un projet arts/sciences initié par l'Université Nice Sophia Antipolis (UNS) et l'association DEL'ART. [1] Ce projet, intitulé *Looking for Search*, a débuté durant l'été 2014. Il a associé en binômes des artistes du groupe FRAME [2] et des scientifiques de laboratoires de l'UNS pour la réalisation d'œuvres autour de la quête d'apprentissage et de la recherche. Les œuvres produites ont été présentées lors d'une exposition en deux temps à l'Avant-Scène, espace d'exposition situé sur le campus Saint Jean d'Angély à Nice, de novembre 2014 à avril 2015. [3]

Dans le cadre de leur collaboration, les deux auteurs de cet article, l'un chercheur en traitement du signal et l'autre plasticien & musicien, ont proposé deux installations, intitulées *Modules de séparation 1 & 2*, représentées en Fig. 1, dans lesquelles le visiteur peut cheminer. Ces œuvres se fondent sur une analogie entre espaces réels et espaces mathématiques (espace signal/données brutes, espace transformée/données traitées) et font plus particulièrement écho aux recherches du premier auteur sur la séparation de sources et la décomposition des signaux. Dans la première installation, en prémisse du projet, des signaux audio mélangés issus des campagnes d'évaluation SiSEC [4] sont diffusés dans le premier espace de la structure tandis que des résultats de séparation (obtenus avec des algorithmes créés par le premier auteur) sont diffusés dans le deuxième espace.

La deuxième installation présente quant à elle le résultat d'un procédé original de synthèse d'un flux visuel guidé par un flux audio (en l'occurrence, musical), dont le contenu de cet article fait une description détaillée. Le flux visuel synthétisé est à la fois *synchrone* et *significatif* du flux audio utilisé en entrée du système ; les deux flux forment ensemble une vidéo cohérente. La méthode de synthèse repose sur une approche apprentissage. Dans une première phase, nous avons sélectionné un ensemble de séquences cinématographiques dont les flux audio et visuel ont fait l'objet d'une décomposition conjointe permettant d'en extraire un ensemble de *motifs audiovisuels*. La structure de ces motifs peut alors être utilisée dans une seconde phase pour générer des images à partir de n'importe quelle entrée audio. Dans le cadre de notre collaboration, nous avons choisi des séquences cinématographiques noir & blanc de quelques secondes mettant à l'écran des *scènes de séparation*, en clin d'œil à l'activité du premier auteur, depuis longtemps dédiée à la *séparation de sources*... Pour ce choix particulier de séquences d'apprentissage, le second auteur a composé une musique ostensiblement larmoyante, produisant au final une séquence audiovisuelle à valeur mélodramatique. Dans notre

1. Association basée à Nice développant des projets dans le domaine de l'art contemporain : http://www.de-lart.org. Un grand merci à Florence Forterre pour son exceptionnel travail de commissaire d'exposition.

2. Groupe composé d'artistes et d'historiens de l'art réunis autour de champs d'investigation communs : http://groupeframe.com

3. Notes d'exposition : http://www.unice.fr/cfevotte/projects/lookingforsearch/catalogue.pdf

4. The Signal Separation Evaluation Campaign : http://sisec.wiki.irisa.fr

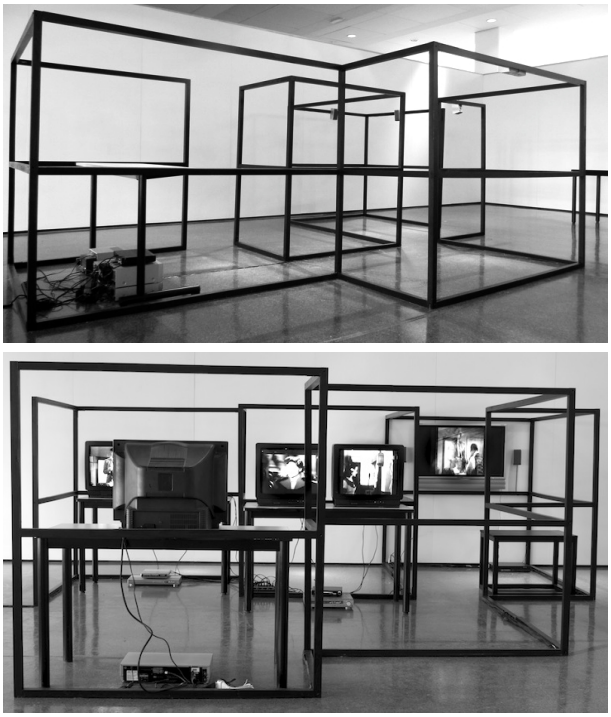


FIGURE 1 – *Modules de séparation 1 & 2*. Dans la première installation, les hauts-parleurs, de petite taille, sont fixés sur la structure (en haut à droite de l'image).

installation, les séquences d'apprentissage sont diffusées (avec le son & l'image) sur des écrans analogiques en entrée de la structure qui mène le visiteur vers un écran numérique plus large diffusant la séquence synthétisée.

La suite de cet article présente dans la partie 2 les détails du procédé d'apprentissage et de synthèse et dans la partie 3 des éléments de résultats obtenus pour l'installation présentée dans le cadre du projet *Looking for Search*.

# 2 Éléments scientifiques

## 2.1 Apprentissage du modèle audiovisuel

### 2.1.1 Représentation des données

Étant donné un ensemble de séquences audiovisuelles d'entraînement, le procédé d'apprentissage, illustré en Fig. 2, est comme suit. Tout d'abord, en l'état actuel de nos travaux et pour des raisons essentiellement de coût de calcul, chaque séquence est traitée individuellement et les motifs audiovisuels appris pour toutes les séquences sont rassemblées au sein d'un même dictionnaire en fin d'apprentissage. Chaque trame visuelle est "vectorisée" puis normalisée (par sa norme $\ell_1$). Le flux audio est segmenté en trames temporelles alignées aux trames visuelles, comme représenté en Fig. 2, dont on retient des descripteurs spectraux Mel (périodogramme ramené à une échelle psychoacoustique). Le vecteur résultant est lui aussi normalisé. Les vecteurs décrivant les trames audio et visuelles sont mis bout à bout dans les colonnes d'une matrice $V$.

### 2.1.2 Factorisation en matrices non-négatives

La matrice $V$ est décomposée par factorisation en matrices non-négatives (NMF), produisant une approximation de rang faible telle que

$$V \approx WH. \tag{1}$$

Les matrices $W$ et $H$ sont à coefficients positifs, contrainte réputée induire une représentation "par parties" [1]. Les colonnes de $W$ forment ainsi des motifs audiovisuels, caractéristiques de la séquence d'entraînement, captant des corrélations singulières entre l'image et le son. Comme illustré en Fig. 2, on note respectivement $W_a$ et $W_i$ les parties audio et visuelles du dictionnaire (indice $a$ comme "audio" et $i$ comme "image"). On note $K$ la dimension commune de $W$ et $H$.

La factorisation (1) est obtenue par minimisation par rapport à $W$ et $H$ de la divergence de Kullback-Leibler de $V$ à $WH$. La minimisation est réalisée sous contrainte de non-négativité des facteurs, mais également sous la contrainte que les colonnes de $W_a$, $W_i$ et de $H$ somment à 1. Dans ces conditions, les colonnes du dictionnaire $W$ et de l'approchée $\hat{V} = WH$ ont la même structure que les données $V$ (parties audio et visuelle chacune normalisées), ce qui nous est apparu comme une contrainte naturelle, bien que non nécessaire. La factorisation peut être obtenue au moyen d'un algorithme majoration-minimisation (MM), voir par exemple [2], procédant à une mise à jour alternée des facteurs et employant des multiplicateurs de Lagrange pour la mise en œuvre des contraintes.

Comme expliqué en début de partie, ce procédé est répété pour chaque séquence d'entraînement, et les dictionnaires $W$ obtenus sont rassemblés dans une matrice que nous appellerons encore $W$. À noter qu'une alternative consisterait à abouter les séquences d'entraînement et à réaliser un apprentissage global sur l'ensemble des séquences. Cette option nécessite cependant des capacités de stockage mémoire au-delà des ressources informatiques à disposition des auteurs.

## 2.2 Synthèse visuelle guidée par l'audio

Une fois la phase d'apprentissage réalisée, un flux visuel peut être aisément synthétisé à partir d'un flux audio arbitraire. Il suffit pour cela de décomposer la représentation temps-fréquence du flux audio, obtenue comme au paragraphe 2.1.1, sur la partie audio $W_a$ du dictionnaire appris. Il en résulte une matrice d'activation $H_s$, qui peut alors être utilisée pour synthétiser un flux visuel $\hat{V}_i = W_i H_s$. Les trames visuelles proprement dites sont obtenues en ré-organisant les colonnes de $\hat{V}_i$ sous forme matricielle et en en projetant les coefficients sur l'intervalle $[0, 255]$ afin d'obtenir une image noir & blanc 8-bits. Le flux visuel synthétisé est par construction aligné temporellement sur le flux audio, et les deux forment ensemble une séquence audiovisuelle cohérente.

Pour un meilleur rendu visuel, la matrice $H_s$ est estimée sous pénalité de lissage temporel de chacune de ses lignes,

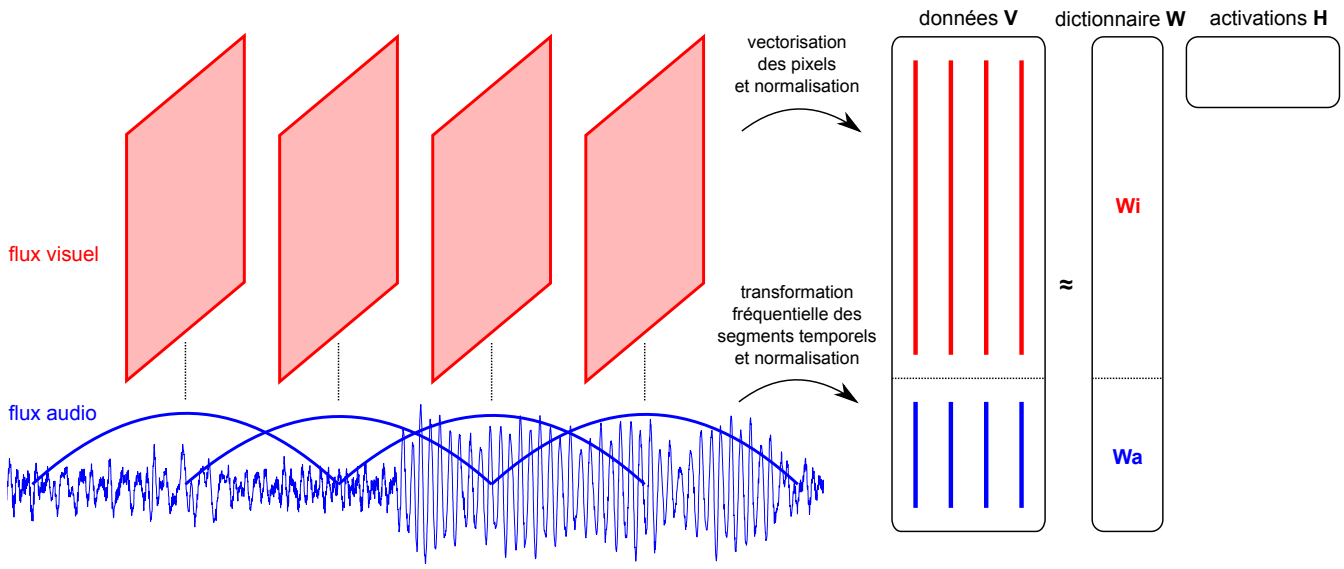



FIGURE 2 – Procédé d'apprentissage des motifs audiovisuels.

en utilisant une variante de l'algorithme présenté dans [3], qui présente une méthode de décomposition non-négative pour la divergence de KL pénalisée par un terme de lissage quadratique.

### 2.3 Relations avec l'état de l'art

Le procédé d'apprentissage présenté est similaire à la technique d'analyse audiovisuelle proposée par Smaragdis & Casey [4]. La construction des données audiovisuelles est en particulier semblable. Leur approche procède toutefois par projection des données et analyse en composantes indépendantes. Dans leur cas la matrice $W$ est obtenue de telle sorte que les lignes de $H = W^T V$ soient le plus mutuellement indépendantes, alors que nous proposons une approche générative qui prend explicitement en compte la non-négativité des données. De nombreuses méthodes de fusion/intégration audiovisuelle ont été proposées en traitement de la parole, pour des tâches de localisation, de reconnaissance, de séparation, e.g., [5, 6]. En particulier, l'approche suivie dans [6] consiste à apprendre des atomes audiovisuels localisés (et invariants) en temps et espace avec une approche codage parcimonieuse ; la partie visuelle des atomes est une courte séquence vidéo (une seule image dans notre cas). L'utilisation faite dans notre travail d'une analyse audiovisuelle pour la synthèse d'un flux d'images est à notre connaissance une démarche nouvelle.

## 3 Résultats d'expériences

### 3.1 Base d'apprentissage

Comme indiqué dans la première partie, nous avons dans le cadre du projet *Looking for Search*, utilisé pour séquences d'apprentissage des extraits cinématographiques noir & blanc (à la fois pour une plus grande facilité de traitement et pour leur valeur esthétique) présentant des scènes de séparation (départ à caractère sentimental d'un protagoniste).[5] Les extraits durent de 9 à 16 secondes et la taille des trames visuelles est de $720 \times 576$ (soit 414.720 pixels) ; la diversité des formats d'image est prise en compte par l'ajout approprié de bandes noires. L'échantillonnage est de 25 trames par seconde. Le son est traité en mono, échantillonné à 22kHz et segmenté en trames de 80 ms (2 fois la période des trames visuelles) dont on extrait 64 descripteurs spectraux Mel.

### 3.2 Résultats d'apprentissage

Les données audiovisuelles de chaque séquence d'apprentissage ont été décomposées en utilisant $K = 3$ motifs. L'algorithme est initialisé avec des matrices non-négatives à coefficients aléatoires. La figure 3 représente pour quelques séquences d'apprentissage les trois motifs audiovisuels appris ainsi que trois trames représentatives des données.

La valeur choisie de $K$ est clairement sous-dimensionnée pour la bonne reconstruction des données (notamment à cause de la grande variabilité des images et l'absence d'invariances dans la modélisation du flux visuel). Cependant, dans le cadre spécifique de ce travail artistique, les motifs visuels appris avec cette valeur nous sont apparus posséder une valeur esthétique satisfaisante et produire des résultats de synthèse plus intéressants qu'avec des décompositions plus fines.

### 3.3 Résultat de synthèse

Une séquence musicale d'environ 10 min a été spécifiquement composée par le second auteur pour le projet *Looking for Search*, séquence utilisée en entrée du

5. Extraits des films *Les Amours d'une blonde*, *Brief Encounter*, *Double Indemnity*, *High Noon*, *Ice Cold in Alex*, *The Misfits*, *Monkey Business*, *The Night of the Hunter*, *La traversée de Paris*, *Zorro's Fighting Legion*.

système de synthèse décrit au paragraphe 2.2. Un extrait de la séquence audiovisuelle produite est disponible à l'adresse `http://www.unice.fr/cfevotte/projects/lookingforsearch`.

# 4 Conclusions

Nous avons décrit dans cet article un procédé d'extraction de motifs audiovisuels dont la structure permet de synthétiser un flux visuel guidé par l'audio. Le procédé repose sur des hypothèse simples, voire simplistes, qui ont cependant permis de produire des résultats intéressants d'un point de vue tout à fait subjectif dans le cadre spécifique d'une création artistique.

Ce projet est voué à se poursuivre, notamment par l'enrichissement de la base d'apprentissage, qui en l'état actuel ne contient que 10 séquences. Compte-tenu du faible nombre de motifs audiovisuels appris pour chaque séquence (seulement 3), notre dictionnaire ne contient pour le moment que 30 éléments, ce qui tend à synthétiser des flux visuels à diversité limitée (d'autant que certains motifs au spectre audio très régulier ont tendance à revenir fréquemment).

Le système lui-même peut être amené à évoluer suivant diverses directions. En particulier, d'autres types de descripteurs pourraient être envisagés pour représenter le flux audio. On pourrait par exemple envisager d'utiliser des descripteurs haut niveau (timbre, genre, etc.) plutôt que bas niveau (i.e., purement spectraux). On pourrait également envisager de calculer ces descripteurs sur des fenêtres d'intégration plus grandes. En revanche, dans la mesure où le système est voué à synthétiser des images, il semble moins aisé d'intégrer des descripteurs plus haut-niveau dans l'analyse de la partie visuelle.

# Références


[1] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 1999.

[2] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 2011.

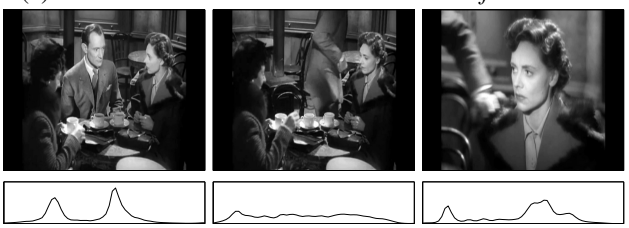
[3] S. Essid and C. Févotte. Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring. *IEEE Trans. Multimedia*, 2013.

[4] P. Smaragdis and M. Casey. Audio/visual independent components. In *Proc. ICA*, 2003.

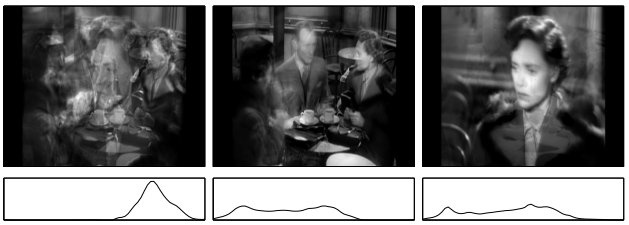
[5] B. Rivet, L. Girin, and C. Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Trans. Audio, Speech and Language Processing*, 2007.

[6] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval. Learning multimodal dictionaries. *IEEE Trans. Image Processing*, 2007.


(a) Trois trames audiovisuelles du film *Brief Encounter*

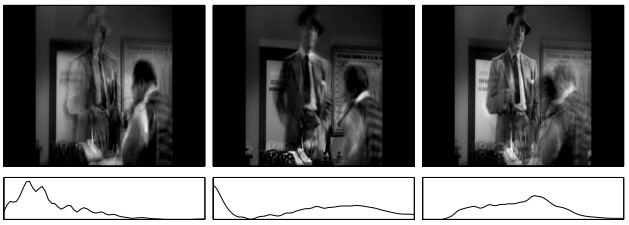


(b) Motifs audiovisuels estimés

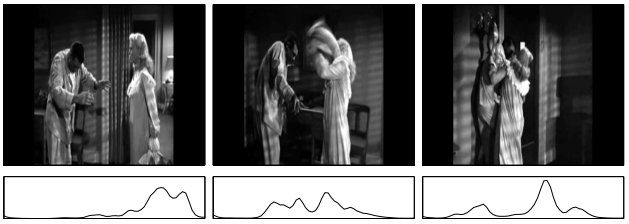


(c) Trois trames audiovisuelles du film *Double Indemnity*

(d) Motifs audiovisuels estimés



(e) Trois trames audiovisuelles du film *Monkey Business*



(f) Motifs audiovisuels estimés

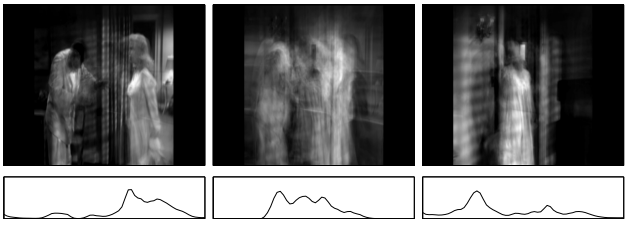


FIGURE 3 – Résultats d'apprentissage. Chacune des images est accompagnée du spectre audio qui lui correspond (échelle Mel couvrant l'intervalle 100-11000 Hz sur 64 bandes).