Nonnegative matrix factorization for unsupervised audiovisual document structuring

Slim Essid[1] and Cédric Févotte[2]

[1] Institut Telecom - Telecom ParisTech, CNRS LTCI
[2] CNRS LTCI, Telecom ParisTech
37 rue Dareau - 75014 Paris, France

July 4, 2011

Abstract

This paper introduces a new paradigm for unsupervised audiovisual document structuring. In this paradigm, a novel Nonnegative Matrix Factorization (NMF) algorithm is applied on histograms of counts (relating to a bag of features representation of the content) to jointly discover latent structuring patterns and their activations in time. Our NMF variant employs the Kullback-Leibler divergence as a cost function and imposes a temporal smoothness constraint to the activations. It is solved for using a majorization-minimization technique. The approach proposed is meant to be generic and is particularly well suited to applications where the structuring patterns may overlap in time. As such, it is evaluated on a person-oriented video structuring task, using a challenging database of political debate videos. Our results outperform reference results obtained by a method using Hidden Markov Models. Further, we show the potential that our general approach has for audio speaker diarization.

Keywords: Content structuring, Unsupervised classification, Machine learning, Videos, Indexing, Bag of features, Matrices.

1 Introduction

Automatic audiovisual document structuring stands as a key technological component as part of the global effort to set up efficient multimedia and video indexing tools. In both the audio and visual domains, highly sophisticated approaches have been proposed in previous works that mostly rely on expert and specific techniques. A number of proposals employ *supervised approaches* exploiting prior knowledge on the general structure of the type of documents to be processed and using domain rules and specific concept or event detectors (typically playing field lines, ball hits and game-related events in sports videos for example) [1, 25]. In our work we are concerned with *unsupervised approaches* that can be applied generically to a wide range of audiovisual documents without the need to assemble training data. In this case, the vast majority of state-of-the art approaches extract the document structure using a form of clustering to group content units that were previously segmented by a change point detection technique. In the video processing domain, these content units are generally shots to be grouped into scenes [30], while in the audio domain they are merely abstract homogeneous content segments (hopefully belonging to different sound classes such as music, silence, speakers, etc.), generally found by a variant of the Bayesian Information Criterion technique [21].

For a wide range of audiovisual documents, for instance news programs, TV talk shows or series, a semantically rich and useful video structure is one that is deduced after onscreen person and/or speaker segmentation [25]. That is a segmentation mapping to the occurrences of onscreen persons and speakers at every instant of the document (which is widely known as *speaker diarization* in the audio field). In this case, the structuring events (here speaker/person occurrences) may overlap in time, hence creating a serious difficulty for classic approaches where each segment of data is assumed to pertain to one of several clusters. Consequently, when multiple events occur in some segments, each possible combination of events should be modeled by a specific cluster. This is a combinatorial approach which may turnout inefficient when the data is scarce.

In this paper we resort to a different approach which explicitly accommodates the composite nature of audio and video data. By composite we refer to the possible simultaneous occurrence of multiple events. First, and like previously mentioned methods, our approach takes the audio or video data (a given file) as a time sequence of frames. In the video case, a frame is simply a single image. In the audio case, a frame is a fixed-length audio segment (1.5 s in this paper experiments) and adjacent frames typically overlap in

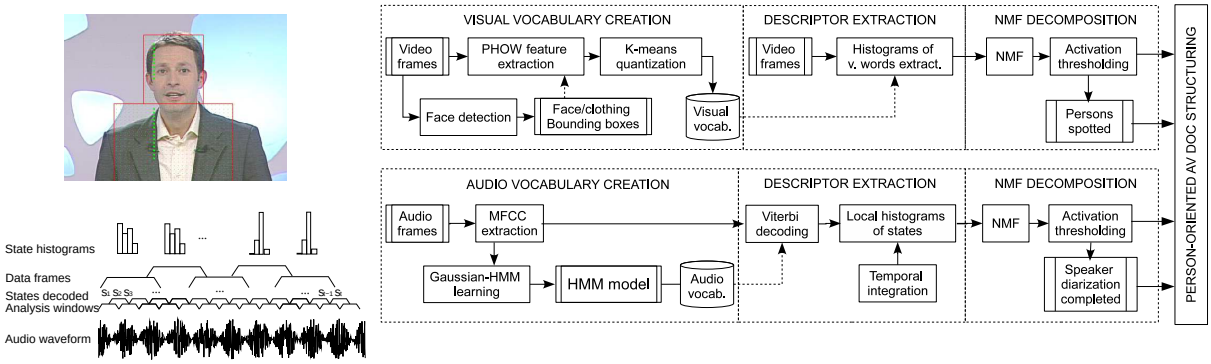

Figure 1: Approach overview. Top: PHOW feature extraction on a visual frame during vocabulary creation phase (cf. Section 5.1). Bottom: Extraction of the state histogram descriptor on an audio signal (cf. Section 5.2).

time. In our approach, each data frame is transformed into a "bag of words", where the term "word" here refers to a local attribute and frames are characterized by occurrence counts of these local attributes (in a analogy with text retrieval, a frame is like a text document characterized by word counts). The set of local attributes, referred to as "vocabulary" is file-specific and learnt for the entire set of frames as later described. Similarly to probabilistic Latent Semantic Indexing (pLSI) [11], or more generally nonnegative matrix factorization (NMF) with the Kullback-Leibler (KL) divergence [9], we propose to factorize the resulting histogram data as the product of a "dictionary" matrix times an "activation" matrix. The columns of the dictionary, akin to "topics", will reflect the individual speaker/person signatures and possibly other components such as image background or audio residual noise. Because time correlation is an important feature of audio and video data, we introduce a novel KL-NMF algorithm that incorporates a smoothness constraint on the activation matrix. Inspired by the work of Ding et al. [5] in the case of NMF with the Euclidean distance, we also introduce a "convex" variant of the KL-NMF algorithm, compatible with the smoothness constraint, which consists in constraining the dictionary elements to be linear combinations of data points. Despite being more computationally intensive than standard NMF, the convex variant will be shown necessary in the audio case, in which the data exhibit less structure.

Generally, the contributions are twofold. First, at the methodological level, we propose a new generic structuring paradigm whereby, whatever the modality (audio or video), NMF is applied on histogram descriptors relating to a bag of features representation, to jointly discover latent structuring elements and their activations in time. Second, at the algorithmic level, we describe a majorization-minimization algorithm for novel smooth and convex variants of KL-NMF.

Note that NMF has been considered for the related task of audio or video classification with diverse usages, but generally at the feature extraction stage. For example, a notorious application of NMF is local feature extraction from face images [13, 17]. In our setting, NMF is instead used at the classification step, after the bag of words transformation. The closest to our work is probably [10] which considers classification of landscape images based on NMF of local color histograms. Our work considerably develops both the feature extraction and factorization parts, and its application to the multimedia segmentation problem is, to our best knowledge, entirely novel.

The outline of the paper is the following. We start by an overview of our approach in Section 2, and present the NMF algorithms in Section 3. We then present two applications of our paradigm in Section 5 that are evaluated in Section 6, before we suggest some conclusions.

2 Approach overview

Our recipe can be roughly accomplished as follows:

1. create a low-level (visual/audio) word vocabulary and use it to extract histograms of word occurrences from the sequence of observation frames at the temporal granularity of interest;

2. apply a variant of Nonnegative Matrix Factorisation (NMF) on the matrix assembled by stacking the word-histogram descriptors column-wise, using the Kullback-Leibler (KL) divergence, adding convexity and temporal smoothing ingredients, so as to extract latent structuring events from the document and their activations across its duration.

Both this general approach to audiovisual document structuring and the NMF variants we propose are completely novel. We will show that NMF is able to discover relevant structuring events as they are

by essence recurrent events. The scheme proposed here is in fact totally generic without preventing one from constraining the semantics of the structure to be extracted. Indeed, the semantics can be imposed by a proper choice of vocabulary. For instance, for the application chosen in this paper to instantiate our paradigm, the features used are relating to audio or visual attributes characterizing speakers or onscreen persons. We believe any other type of structures could be extracted following the same scheme merely by adapting the features and the observation time horizons.

An overview of our approach applied to person-oriented video structuring is depicted in Figure 1. The visual and audio streams are first processed in two different threads following the same recipe mentioned earlier. The NMF algorithm represents the (audio or visual) word-histogram descriptors as the activations of particular basis vectors (to be associated in this application with target persons) at every time instant. By thresholding the activations, speaker/person spotting information is deduced.

The vocabulary words are extracted specifically for the video being processed following different procedures for the audio and visual modalities. For our person-oriented structuring tasks we adopted the following:

- For the visual content, the vocabulary is constructed using only the frames where faces have been detected. A variant of dense SIFT features is extracted from face and clothing regions of the image and used to build the visual word dictionary (which will be further described in Section 5.1).
- For the audio, a low-level segmentation is obtained by fitting a Hidden Markov Model (HMM) to the sequence of short-term Mel Frequency Cepstral Coefficients (MFCCs) extracted from the signal, before Viterbi decoding, hence labeling the audio frames with the decoded sequence of states. The audio word vocabulary is thus composed of the set of HMM states learned, and a bag of features representation is obtained by counting the occurrences of states in a temporal integration window covering a sequence of local frames (see Section 5.2 for details).

Once the descriptors have been extracted, they are processed by an NMF algorithm, as explained in the next section.

3 Smooth and convex NMF for histogram sequences processing

3.1 Motivation

Given histogram data V with coefficients v_{fn} representing the contribution of descriptor-component f at frame n, we seek a factorization of the form

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

where W and H are nonnegative matrices of dimensions $F \times K$ and $K \times N$, respectively, with coefficients w_{fk} and h_{kn}. We will denote by v_n, w_k and h_n the columns of V, W and H, respectively. We seek to retrieve patterns characteristic of each individual speaker/person in the columns of W while the rows of H represent the activation of these patterns along the video. Because we assume an additive model in the data domain, we allow two speakers to be active in a same frame n. This is in contrast with usual mixture of distributions models which instead assume a model of the form $v_n \approx h_{kn} w_k$ with probability α_k, i.e., a model in which each data frame v_n is the expression of a unique "event" (either a single speaker, or a certain combination of speakers, but where each possible combination has to be modeled by a specific state). Given a factorization of the form (1) we will base our speaker detection criterion on the amplitudes of the coefficients of H, using appropriate thresholding.

3.2 Specifications

3.2.1 Measure of fit

We seek an approximate factorization (1) in the Kullback-Leibler (KL) sense, i.e., such that $D_{KL}(V|WH)$ is small, where

$$D_{KL}(V|WH) = \sum_{fn} d_{KL}(v_{fn}|\sum_{k} w_{fk}h_{kn}) \tag{2}$$

and where

$$d_{KL}(x|y) = x \log \frac{x}{y} - x + y \tag{3}$$

is the generalized KL divergence (sometimes referred to as I-divergence). The generalized KL divergence is commonly used measure of fit for histogram data, and in particular, in the context of NMF, it derives from a natural probabilistic model, see, e.g., [9].

3.2.2 Smoothness

Because we are dealing with time series of histograms, a certain amount of correlation is to be expected between columns of H. As such, we propose to regularize the factorization (1) by a smoothness-favoring penalty on H, chosen as

$$S(H) = \frac{1}{2}\sum_{k=1}^{K}\sum_{n=2}^{N}(h_{kn} - h_{k(n-1)})^2 \tag{4}$$

More elaborate smoothness constraints, derived in a Bayesian setting from hierarchical Gamma chains, and offering a shape tuning parameter, have also been considered in the audio literature [29], but we here resort to the more standard smoothness measure (4) for which we will derive an original algorithm in Section 4.1.

3.3 Forming the objective function

Assembling the previous specifications, we are left with the following minimization problem:

$$\min_{W,H} C(W,H) \stackrel{\text{def}}{=} D_{KL}(V|WH) + \beta S(H) \quad s.t \quad W \geq 0,\, H \geq 0 \tag{5}$$

where β is a fixed nonnegative scalar, weighting the penalty, and $A \geq 0$ expresses nonnegativity of the coefficients of matrix A. As it turns out, a solution $(W^\star, H^\star)$ to (5) may only satisfy $\|W^\star\| \to \infty$ or $S(H^\star) = 0$ (i.e., $h^\star_{kn}$ is constant w.r.t n). To see this, let us assume that there exists a solution to (5) such that $\|W^\star\| < \infty$ and $S(H^\star) \neq 0$. Let Λ be a diagonal matrix of "scale" factors λ_k, with $0 < \lambda_k < 1$, and let $W^\bullet = W^\star\Lambda^{-1}$, $H^\bullet = \Lambda H^\star$. It follows $C(W^\bullet, H^\bullet) = D_{KL}(V|W^\star H^\star) + \beta\sum_k \lambda_k^2 S(\underline{h}^\star_k)$, where $\underline{h}_k$ denotes k^{th} row of H. Thus, we obtain $C(W^\bullet, H^\bullet) < C(W^\star, H^\star)$, i.e., a contradiction. As such it appears necessary to control the norm of W, and we propose to subject the minimization (5) to the additional constraint that $\|w_k\| = 1$, where $\|\cdot\|$ is taken in the following as the ℓ_1-norm. When $S(H^\star) \neq 0$, this prevents from $\|W^\star\| \to \infty$ and when $S(H^\star) = 0$ (an unlikely but admissible solution) this simply solves the scale indeterminacy that exists between W and H. In the end, we want to solve

$$\min_{W,H} C(W,H) = D_{KL}(V|WH) + \beta S(H) \quad s.t \quad W \geq 0,\, \|w_k\| = 1,\, H \geq 0 \tag{6}$$

As it appears, and following [6, 15], the minimization (6) is equivalent to the minimization of the following scale-invariant objective function:

$$\min_{W,H} \bar{C}(W,H) \stackrel{\text{def}}{=} D_{KL}(V|WH) + \beta\, S(\Lambda H) \quad s.t \quad W \geq 0,\, H \geq 0 \tag{7}$$

where $\Lambda = \text{diag}\,(\|w_1\|, \ldots, \|w_K\|)$. Indeed, let (W, H) be a pair of nonnegative matrices and let $(W^\bullet = W\Lambda^{-1}$, $H^\bullet = \Lambda H$ be their rescaled equivalents. Then, we have $\bar{C}(W,H) = C(W^\bullet, H^\bullet)$, and $W^\bullet$ satisfies the constraint $\|w^\bullet_k\| = 1$ by construction. As such, one may solve (7), free of scale constraint, and then rescale its solution to obtain a solution to (6). We will use the notation $\lambda_k = \|w_k\|$ in the rest of the paper. The next section describes a majorization-minimization (MM) algorithm for the resolution of (7).

4 Majorization-minimization for smooth and convex KL-NMF

We describe an iterative algorithm that updates H given the current iterate of W and then W given the current iterate of H. Our algorithm employs no heuristics and is derived in a rigorous maximisation-minimisation framework, which guarantees non-increaseness of the objective function at each iteration. Sections 4.1 and 4.2 describe the updates of H and W, respectively.

4.1 Update of H given W

4.1.1 Unpenalized case ($\beta = 0$)

In the unpenalized case and given W we are left with

$$\min_H \; C(H) = D_{KL}(V|WH) = \sum_n D_{KL}(v_n|Wh_n) \quad s.t \quad H \geq 0. \tag{8}$$

Because the objective function separates into independent contributions of h_n, $n = 1, \ldots, N$, we are essentially left with the problem of minimizing of $C(h_n) = D_{KL}(v_n|Wh_n)$. This is a standard nonnegative

linear regression problem which may be handled in a majorization-minimization (MM) framework [12], based on the iterative minimization of an (easier to minimize) auxiliary majorizing function. The $\mathbb{R}_+^K \times \mathbb{R}_+^K \to \mathbb{R}_+$ mapping $G(h|\tilde{h})$ is said to be an *auxiliary function* to $C(h)$ if and only if 1) $\forall h \in \mathbb{R}_+^K$, $C(h) = G(h|h)$, and 2) $\forall (h, \tilde{h}) \in \mathbb{R}_+^K \times \mathbb{R}_+^K$, $C(h) \leq G(h|\tilde{h})$. The optimization of $C(h)$ can be replaced by iterative optimization of $G(h|\tilde{h})$. Indeed, any iterate $h^{(i+1)}$ satisfying $G(h^{(i+1)}|h^{(i)}) \leq G(h^{(i)}|h^{(i)})$ produces a monotone algorithm (i.e., an algorithm which decreases the objective function at every iteration) as we have $C(h^{(i+1)}) \leq G(h^{(i+1)}|h^{(i)}) \leq G(h^{(i)}|h^{(i)}) = C(h^{(i)})$. As described in [4, 14, 8], an auxiliary function $G(h_n|\tilde{h}_n)$ to $C(h_n)$ can be constructed using Jensen's inequality thanks to convexity of $C(h_n)$, leading to

$$G(h_n|\tilde{h}_n) = \sum_k -\psi_{kn} \log h_{kn} + \lambda_k h_{kn} + cst, \tag{9}$$

where $\psi_{kn} = \tilde{h}_{kn} \sum_f w_{fk} v_{fn} / \tilde{v}_{fn}$, with $\tilde{v}_{fn} = \sum_k w_{fk} \tilde{h}_{kn}$, and *cst* denotes constant terms w.r.t $\tilde{h}_n$. The minimization of $G(h_n|\tilde{h}_n)$ w.r.t $\tilde{h}_n$ leads to the standard multiplicative update $h_{kn} = \psi_{kn}/\lambda_k$

4.1.2 Penalized case ($\beta > 0$)

In the penalized problem, the contribution of h_n to $\bar{C}(H) = D_{KL}(V|WH) + \beta S(\Lambda H)$, $1 < n < N$, can be written as

$$\bar{C}(h_n) = D_{KL}(v_n|Wh_n) + \beta\, L(h_n; h_{n-1}, h_{n+1}), \tag{10}$$

where

$$\begin{aligned} L(h_n; h_{n-1}, h_{n+1}) &= \frac{1}{2} \sum_k \lambda_k^2 \left[(h_{k(n+1)} - h_{kn})^2 + (h_{kn} - h_{k(n-1)})^2 \right] \\ &= \beta \sum_k \lambda_k^2 \left[h_{kn}^2 - (h_{k(n+1)} + h_{k(n-1)}) h_{kn}) \right] + cst, \end{aligned}$$

and where *cst* is a constant of h_{kn}. Using the preceding results, an auxiliary function to the penalized objective function $\bar{C}(h_n)$ is readily obtained as

$$G_\beta(h_n|\tilde{h}_n) = G(h_n|\tilde{h}_n) + \beta\, L(h_n; h_{n-1}, h_{n+1}). \tag{11}$$

The minimization of $G_\beta(h_n|\tilde{h}_n)$ for $1 < n < N$ is easily shown to amount to solving an order 2 polynomial with a single positive root, given by

$$h_{kn} = \frac{\sqrt{b_{kn}^2 + 4 a_{kn} \psi_{kn}} - b_{kn}}{2 a_{kn}}, \tag{12}$$

where $a_{kn} = 2\beta\lambda_k^2$, $b_k = \lambda_k(1 - \beta\lambda_k(h_{k(n-1)} + h_{k(n+1)}))$, $1 < n < N$. At the border of the chain, $n = \{1, N\}$, the penalty (11) reduces to only one of its two terms and we obtain $a_{k1} = \beta\lambda_k^2$, $b_{k1} = \lambda_k(1 - \beta\lambda_k h_{k2})$, and $a_{kN} = \beta\lambda_k^2$, $b_{kN} = \lambda_k(1 - \beta\lambda_k h_{k(N-1)})$.

In practice, given $\tilde{V} = W\tilde{H}$ (with coefficients $\tilde{v}_{fn}$ on which ψ_{kn} depends) computed from current iterate $\tilde{H}$, the columns h_n of H are updated iteratively with replacement for $n = 1, \ldots, N$ using (12). $\tilde{V}$ is then updated with the new value $\tilde{H} = H$, and the algorithm proceeds to next iteration.

4.2 Update of W given H

4.2.1 Unpenalized case ($\beta = 0$)

In the unpenalized case and given H, we are left with

$$\min_W C(W) = D_{KL}(V|WH) \quad s.t \quad W \geq 0. \tag{13}$$

which is essentially the same problem as (8). As such a suitable auxiliary function for $C(W)$ is

$$G(W|\tilde{W}) = \sum_{fk} -\phi_{fk} \log w_{fk} + \sigma_k w_{fk} + cst, \tag{14}$$

where $\phi_{fk} = \tilde{w}_{fk} \sum_n [v_{fn}/\tilde{v}_{fn}] h_{kn}$ and $\sigma_k = \sum_n h_{kn}$, and one obtains the multiplicative update $w_{fk} = \phi_{fk}/\sigma_k$.

4.2.2 Penalized case ($\beta > 0$)

In the penalized case ($\beta > 0$), we have to solve

$$\min_W \bar{C}(W) = D_{KL}(V|WH) + \frac{\beta}{2} \sum_k s_k \lambda_k^2 \quad s.t \quad W \geq 0, \tag{15}$$

where $s_k = 2S(\underline{h}_k)$ and where we recall that $\lambda_k = \sum_f w_{fk}$ is a function of W. As before, an auxiliary function to the penalized objective function $\bar{C}(W)$ is given by

$$G_\beta(W|\tilde{W}) = G(W|\tilde{W}) + \frac{\beta}{2} \sum_k s_k \lambda_k^2 \tag{16}$$

and the minimization of $G_\beta(W|\tilde{W})$ is easily shown to amount to solving an order 2 polynomial with a single positive root, given by

$$w_{fk} = \frac{\sqrt{b_{fk}^2 + 4a_{fk}\phi_{fk}} - b_{fk}}{2a_{fk}}, \tag{17}$$

where $a_{fk} = \beta s_k$, $b_{fk} = \sigma_k + \beta s_k \sum_{g \neq f} w_{gk}$.

5 Application to person/speaker-oriented video structuring

For a variety of TV shows, a structuring scheme centered on show-participants' occurrences and interventions is particularly meaningful and useful [25]. This is especially true as the distinction between *onscreen person* and *speaker* is made, where the former refers to a person appearing on the current visual frame without necessarily speaking, while the latter refers to a person speaking without necessarily being onscreen. Then, structuring an audiovisual document based on persons/speakers localization essentially boils down to performing two key sub-tasks: *onscreen person spotting* and *speaker diarization*.

Not only are these temporal segmentations useful *per se*, but also they define various higher-level structuring units. Indeed, they constitute relevant entry points to the show content, enabling various navigation modes. For instance, the following: "*browse over all interventions of participant Jack, with Jack speaking and onscreen*", which is typically the type of video segments that would be used to build a summary of Jack's interventions. Further, person/speaker temporal segmentations can be easily used as the basis to deduce a structure such as the one that will be targeted with our evaluation database, where shots are to be categorized into *single participant*, *multiple participants* and *overall*. They are also key for social signal processing applications [27].

5.1 Visual Word extraction for onscreen person spotting

Visual person spotting has been considered in a number of studies [22, 7]. The classic approach consists in detecting faces and using a clustering method on low-level features, the whole process being possibly guided by a shot change detector and a face tracking module. Features used in this context were extracted from the face and possibly the clothing regions, including color, texture and SIFT-like features.

In our work we use a bag of visual words representation based on PHOW features, where PHOW refers to *Pyramid Histograms Of visual Words*. Note that the term *Word* in the acronym PHOW is kept here only to be consistent with the original references [2, 26] where it refers to bins of Histograms of Orientation Gradients (HOG), and should not to be confused with our usage of *visual word* relating to the vocabulary obtained by quantization of the whole set of PHOW features. To avoid confusions, we will use *feature* to refer to the low-level attributes (i.e., PHOW features). The *features* are quantized to create the vocabulary that is used to extract histograms of word occurrences, which will be referred to as *descriptors*.

During the dictionary construction phase, the PHOW features are extracted only from onscreen persons' faces and clothing regions as depicted in the top-left corner of Figure 1. These regions are spotted as follows. First a Viola & Jones face detector [28] is applied on the video frames. Then the clothing region is detected by creating a rectangular bounding box below the face bounding box, similarly to [7]. Its width and height are respectively chosen to be twice and 2.5 times the width and height of the latter. These parameters have been chosen to limit the situations where a part of the background is included in the clothing bounding box.[1]

[1] We will see in Section 6.2 that this constraint does not need to be too rigid, as it is useful for our task to have some visual words representing the background.

It is worth mentioning that though color histograms seem to be natural descriptors of the clothing regions [7, 24], we found them to be less reliable for our task than the descriptors we propose. In fact, we performed extensive preliminary testing with a number of color histogram variants (testing different color spaces and quantization steps) and found them to be systematically lacking robustness to the significant illumination changes accompanying camera viewpoint changes in the talk show videos used for our evaluation.

PHOW features are extracted on a 8-pixel step grid at 3 scales using bin sizes of 8, 16 and 32-bins [26]. The set of all PHOW features extracted from regions of interest over all frames of the current video where a face has been detected, are then quantized on 128 bins using the K-means algorithm. All parameters have been tuned once and for all on a development video that will not be included in the evaluation, to test for the generalization ability of our system.

The visual vocabulary thus obtained (specifically for the current video) is used to extract histograms of word counts from every frame of the video. Face detection is no longer used at this stage, that is PHOW features are extracted over the whole frames, which are thus globally described by the histograms of visual words, allowing us to cope with the face detector misses, especially on wide shots.

Therefore, we are relying on the NMF algorithm to decompose global frame-based histograms of words, possibly representing the joint occurrence of two or more persons, into elementary histograms, each representing a single person. Note that, the process is clearly facilitated by the fact that there are numerous close-up shots in a TV program video, showing only one person at a time.

As previously explained, only the descriptors need to be adapted to each particular task, the rest of the temporal segmentation scheme remaining generic.

5.2 Audio descriptor extraction for speaker diarization

For audio analysis the temporal evolution of the local signal characteristics is of great importance. This has led researchers in the field to largely rely on dynamic modeling approaches, hence the success of HMMs for audio classification tasks in general, and in particular for speaker diarization tasks where it is used with Gaussian Mixture Model (GMM) emission probabilities (see for example [19]). In fact, agglomerative clustering techniques exploiting GMM-HMM structures and Binary Information Criteria over cepstral features have been extensively used as it has proven successful in solving this problem (for instance within NIST[2] international evaluation campaigns).

HMMs are traditionally used as a decision model in the sense that a one-to-one mapping is determined between the speakers and the hidden states, and the diarization result is directly deduced by Viterbi decoding of the observed sequence of low-level features (generally MFCC features) being modeled by the GMM-HMM. In this work, we follow a different approach, inspired by [16], where we use HMMs only to build the audio descriptors and leave the speaker modeling and decision taking tasks to the NMF algorithm. The bottom-left corner of Figure 1 sums up the whole descriptor extraction procedure. The audio signal is analyzed in short overlapping 20-ms length windows, with a 10-ms hop size, over which 12 Mel Frequency Cepstral Coefficients are extracted (excluding the energy coefficient). A Q-state HMM is trained in a non-supervised fashion on the sequence of MFCCs, with Q much greater than the expected number of speakers, using Gaussian state-conditional densities with full covariance matrices. The audio word vocabulary merely consists of the HMM states found by the Baum-Welch learning algorithm. The most likely sequence of states is then inferred by Viterbi decoding yielding a state-label for each low-level frame. Subsequently, state occurrences are counted over 1.5-s length integration windows using a 40-ms hop size, hence forming the audio descriptors (extracted at a rate of 25 Hz).

6 Experimental validation

Hereafter, a thorough evaluation of a visual-only person-oriented structuring system is first presented on a challenging database of political debates videos. Subsequently, we propose a second instantiation of our generic structuring scheme on a speaker diarization problem. We merely aim to make a proof of concept (on a single video) for the latter, in order to emphasize that our approach can be truly applied to various tasks, and to show its potential for complex problems, such as speaker diarization, where it exhibits its capacity to cope with *overlapped speech* segments, an issue that remains critical for researchers in this field [31].

[2]National Institute of Standards and Technology: http://www.nist.gov/index.html

6.1 Evaluation database

We exploit the *Canal9 political debates* database for our evaluation [27]. This is a challenging TV show database meant to serve for research on automatic analysis of social interactions. It covers 4 years of broadcast. Each broadcast features a moderator and 2 to 4 guests debating a political question. There are different guests from show to show and both the moderator and the set may vary, though most of them have been shot in the same studio set.

The database comes with different types of manual annotations. The visual annotations define an interesting structuring scheme based on a particular taxonomy of the shots relating to camera viewpoints, which is illustrated in Figure 2. Every shot has been classified into one of three categories, namely *"full group"*, *"multiple participants"* and *"personal shot"*. Additionally, manual identification of the participant appearing onscreen is given on "personal shots".

In order to assess the robustness of our system, we use in our evaluation 10-minute video excerpts from each of the 41 first shows,[3] hence exploiting around 7 hours of video content, involving 189 distinct persons and totaling 28521 video shots. All system parameters tuning has been done once and for all on a single development video excerpt (labeled 06-11-22 in the database) leaving 40 videos for the evaluation. This procedure is meant to show that our system is able to generalize properly despite the limited tuning effort.

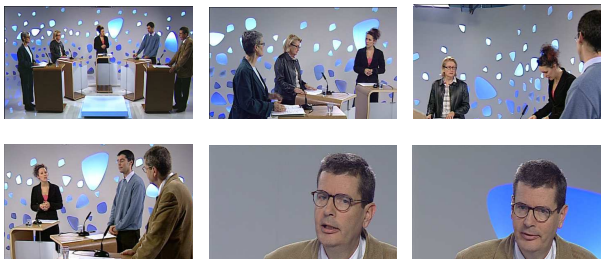

Figure 2: Canal9 annotated shot-types. First shot (upper-left image) is labeled *"full group"*, next shots: *"multiple participants"*, and last 2 shots are *"personal shots"* labeled with the identity of the onscreen person.

The database is quite challenging as most camera viewpoints are not stable in time, even across shots depicting the same set of participants (as can be seen in Figure 2), which is also accompanied with significant changes in illumination. The "full group" shots are an exception to this, though, as they repeat invariably over the show duration.

6.2 Visual-based structuring evaluation

Our visual system tries to automatically replicate the Canal9 database visual groundtruth structure in a non-supervised fashion, hence without trying to assign the given shot labels, or to name the participants on the personal shots. Rather we aim at jointly clustering the shots of the same category and the "personal shots" of the same participants. This indeed defines a semantically meaningful person-oriented structuring scheme since the different shot changes and viewpoints implicitly translate a high-level human structuring process, that is the one proposed by the TV show director who generally selects for the viewer the viewpoints that are the most informative about the participants' interventions and reactions.

6.2.1 Reference system and evaluation procedure

A reference system has been implemented that uses HMMs to model the same sequence of visual-word histograms exploited by our NMF system. These HMMs employ multivariate Gaussian emission probabilities with full covariance matrices. The number of hidden states is set to $N_{sp} + 2$, where N_{sp} is the number of current-show participants. $N_{sp} + 2$ is exactly the number of target categories: one for the "full group" shots, one for the "multiple participants" shots, and one for each participant's "personal shots".

Thus we suppose the number of participants to be known, both for the reference system and our NMF-based system, which is often acceptable as it can be deduced from textual metadata attached to the TV content (typically integrated subtitles and/or teletext, see for instance [22]), or given by an operator in human-assisted systems. Alternatively, model order selection techniques could be employed which has proven successful especially in the NMF case [23].

[3] Excluding the pilot show labeled 05-09-21, for which the groundtruth annotation is missing.

Figure 3: KL-NMF activations, i.e., H coefficients on a short excerpt of the development video with $\beta = 0$. Each subplot represents the temporal sequence of activations for one w_k component, $1 \leq k \leq K = N_{sp}+1$. For each component, the image on the right corresponds to the frame where the activation value is maximum, which is supposed to be a good representative of the content modeled by the corresponding w_k component. Red vertical lines are groundtruth shot boundaries and the other images inside the plot or around it are key frames of the time-corresponding shots. Green dotted horizontal lines are decision thresholds. It can be seen that NMF has succeeded in extracting the relevant components and related activations. Note that the 2nd component is not activated here as the corresponding person does not appear in any personal shot of this part of the video.

Scoring is performed following NIST speaker diarization evaluation procedure[4] [20] which is well adapted to our problem. It consists in finding a one-to-one mapping between groundtruth segment labels (here shot types and person identities on "personal shots") and the labels found automatically for each segment of the video, such that the total time that is shared between the groundtruth labels and the corresponding system outputs is maximized over the whole show duration. This is done with the constraint that each reference label be mapped to at most one system output label. As suggested by the NIST procedure, 0.25 s time collars are used on the segment-boundaries to forgive potential errors in the groundtruth.

The evaluation metric is thus the overall shot-type based segmentation error. Note that we are evaluating our high-level person-oriented structuring task, rather than an on-screen person spotting task. We unfortunately cannot accurately evaluate the latter since the groundtruth does not indicate who the onscreen-persons are on the "full group" and "multiple participant" shots. It is worth mentioning, though, that in our observations the NMF-based systems seem to behave well even for this low-level task.

6.2.2 Analysis of the NMF output

NMF is computed using $K = N_{sp} + 1$ components. This choice has been made to let the NMF algorithm extract one histogram component for each person, plus one for the histogram-descriptor observations which are dominated by visual words describing the background. "full group" frames are an example of such observations that are systematically captured by one NMF component as can be seen in Figure 3. Clearly, this type of shots are easily represented by our method due to their highly stable and recurrent nature. From this Figure, it can also be noted that there are lower amplitude activations on this same component, that relate to "multiple participant" shot occurrences. These amplitudes are lower since fewer elements of the studio background appear on the corresponding tighter shots (and actually even

[4] We actually use the NIST scoring scripts.

fewer on "personal shots" causing the current component not to be activated for the latter). In fact, occurrences of background-related visual words are initially highly present on all observations, which is why all histogram vectors are normalized (prior to NMF computation) by dividing each row of matrix V by the row maximum value, so that each descriptor coefficient have full dynamics and the cost is not dominated by histogram bins with large amplitudes (thus typically bins relating to background visual words). One might wonder how come such visual words are present in the vocabulary while it was learned from features extracted in persons' face and clothing bounding boxes. Recall, though, that we intentionally did not try to be too rigid on the location of these bounding boxes, hence allowing us to capture some elements of the background as can be seen in Figure 1. Additionally, background elements are unintentionally captured on every "false-alarm face detection", which here is useful to our system, the key idea being that we mostly want visual words representing the onscreen persons, but also a few to describe the background.

The desired video structure is obtained by thresholding the activations (see Figure 3). The thresholds are chosen (once and for all on the development video) to be 0.6 times the maximum activation value for each component. This yields $N_{sp} + 1$ clusters (one cluster per NMF component) covering the N_{sp} speakers and the "full group" frames as can be deduced from Figure 3. A frame belongs to a cluster if its corresponding activation is above the decision threshold. A last cluster is created with all unassigned segments which are associated to situations where all corresponding activations are below the chosen threshold. This is always a winning strategy (as will be confirmed by the results on the whole database), thanks to the behavior of the "background-related" component (top first component in Figure 3), where as previously explained two levels of activations are observed: one corresponding to the "full group" shots and the other to the "multiple participant" shots. It is important to note that none of our systems exploit a shot change detection module. Instead shot boundary detection comes as natural byproduct of our higher-level structuring process. In fact, both the reference HMM system and our NMF-based system prove very successful at detecting shot changes.

Figure 4 illustrates the effect of the smoothing on the h_{kn} sequences. The activations become more stable and easier to threshold, hence potentially creating a positive impact on the system performance (as will be seen in the next sections).

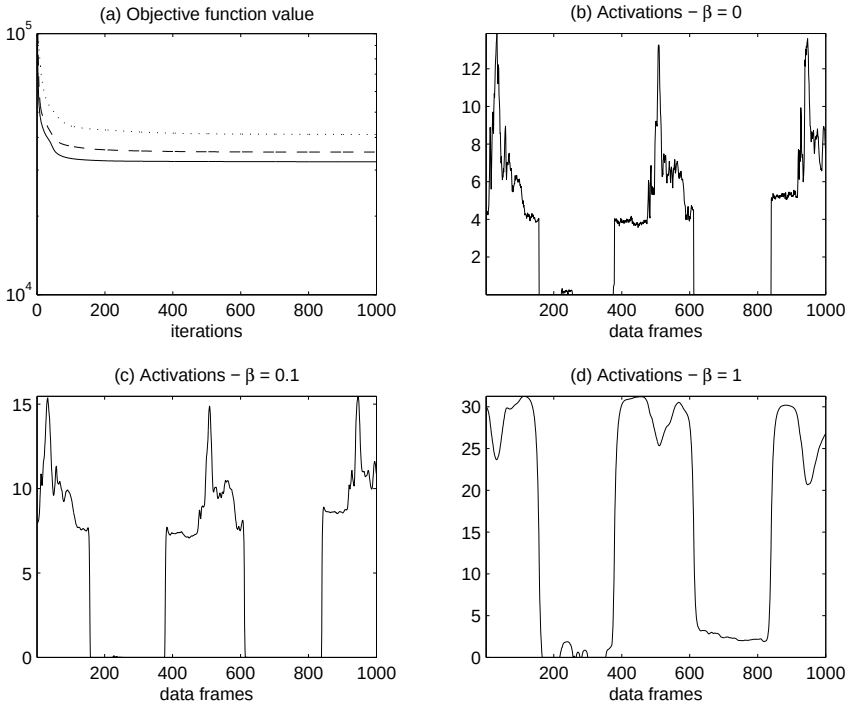

Figure 4: Smooth KL-NMF results on video 06-10-04 (visual stream); $F = 128$, $N = 15001$ and $K = 6$. (a) Cost functions for $\beta = 0$ (solid line), $\beta = 0.1$ (dashed), $\beta = 1$ (dotted). (b-d) First 1000 coefficients of $\underline{h}_1$ obtained with the three values of β. One thousand iterations of the unpenalised and penalised algorithms take respectively 349 and 362 seconds with a MATLAB implementation on a 2.8 GHz Quad-Core Mac with 8 GB RAM.

6.2.3 Evaluation results discussion

The overall structuring errors of our NMF-based systems are 16.6%, 14.6% and 26.2%, respectively for $\beta = 0, 0.1$ and 1. The overall performance of the HMM reference system is 23.8%. The statistics of these scores across all database videos are summed-up in the boxplots of Figure 5.

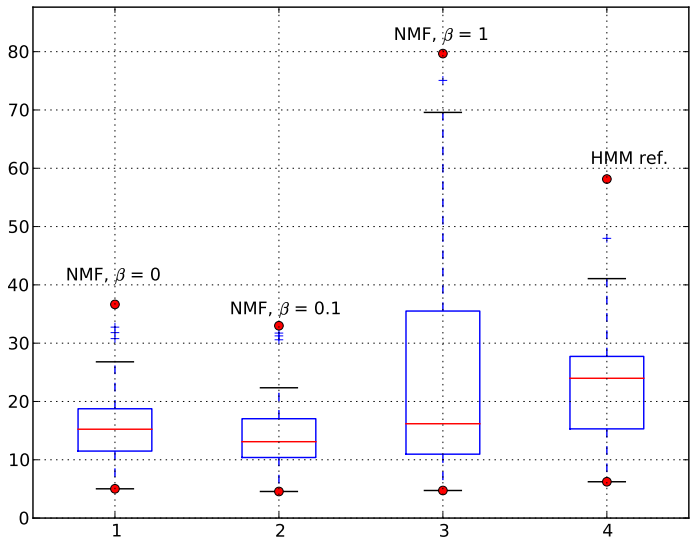

Figure 5: Overall visual structuring error in % and box plots of the per-show visual structuring errors in %. Whiskers extend to the most extreme scores within 1.5 times the inner-quartile range.

NMF-based systems are clearly superior to the reference HMM system with $\beta \in \{0, 0.1\}$. The error can be as low as 4.6% with NMF(0.1) against 6.4% with HMMs, and never exceeds 33.4% for the former while it may be as high as 58.3% for the latter.

Further, there is a significant improvement with the smooth NMF version ($\beta = 0.1$) compared to the non-smooth "standard" version ($\beta = 0$), with -2% in absolute error. This is no longer true if too much smoothing is imposed, as asserted by the poor results obtained with $\beta = 1$, where a too strong smoothing penalty may have negtively affected the extraction of the relevant basis vectors.

6.3 Speaker diarization system

We have shown in the visual stream segmentation example that very competitive results can be obtained with a "standard" KL-NMF approach with no specific assumed structure for W (and with the possible additional smoothness penalty on H). In other examples relying on less structured data, such as audio segmentation, we observed that the standard NMF approach may fail at extracting single speakers individual patterns and may instead extract elementary "parts" of speakers, possibly shared among several speakers. This is a known property of NMF [13], which can be desirable in some settings, such as coding, but not in ours. As such, it can be beneficial to assume a particular structure on W that penalizes the latter effect. In our setting, though multiple speakers occur in many frames, each speaker is also expected to appear alone in a large proportion of data (corresponding to single speaker segments in the audio track). Hence, the individual speaker patterns may be retrieved from the data itself and we may assume the dictionary matrix W to be a linear combination of data points, i.e., $W = VL$, where L is a nonnegative $N \times K$ "labeling" matrix. This corresponds to a "convex"-NMF setting, as proposed by Ding et al. [5], where the authors show that the matrix columns of L tend to become sparse, i.e., the columns of W are built from a linear combination of a few data points, acting as "centroids". Ding *et al.* [5] consider convex-NMF with the Euclidean distance, but we obtained similar findings with the KL divergence. It is possible to combine the results of [8], which reports MM updates for convex-NMF with the KL divergence, with the results of Section 4.2 to produce a MM algorithm for smooth & convex KL-NMF. Given $W = VL$, the update of H given by Eq. (12) is unchanged.

A suitable MM update for L, with coefficients l_{mk}, can be obtained as

$$l_{mk} = \frac{\sqrt{b_{mk}^2 + 4a_{mk}\phi_{mk}} - b_{mk}}{2a_{mk}}, \tag{18}$$

where $\phi_{mk} = \tilde{l}_{mk} \sum_{fn} v_{fm}[v_{fn}/\tilde{v}_{fn}]h_{kn}$, $a_{mk} = \beta s_k \delta_m^2$, $b_{mk} = (\sigma_k + \beta s_k \sum_{n \neq m} \delta_n l_{nk})\delta_m$, and $\delta_m = \sum_f v_{fm}$. It has to be noted that the update of W (i.e., L) in convex NMF is of complexity $\mathcal{O}(N^2K)$ (per iteration) while of complexity $\mathcal{O}(FNK)$ in standard NMF. Given that in our setting $N >> F$, convex NMF induces an important increase of the computational burden.

Figure 6 depicts the activations found by our convex NMF algorithm, with $\beta = 0.5$, applied to the audio-word histograms of our development video. $Q = 80$ states were used in the HMM exploited for the histogram descriptor extraction. The result is quite promising as the activations presented in this Figure are easy to threshold and are found to faithfully represent the groundtruth, in the sense that each NMF component effectively represents a different speaker with the appropriate activations.

We believe our method has a great potential for this task, chiefly for its ability to cope with overlapped speech segments as can be observed around time $t = 9000$ frames, where this situation occurs. Two components are then active, that correspond to the two persons who are effectively speaking simultaneously at that instant.

It is important to note that it was necessary to use both the smoothing and convexity ingredients to get these results. The non-convex NMF version did not behave well as it tended to decompose a same speaker on two different components and to represent others with the same component.

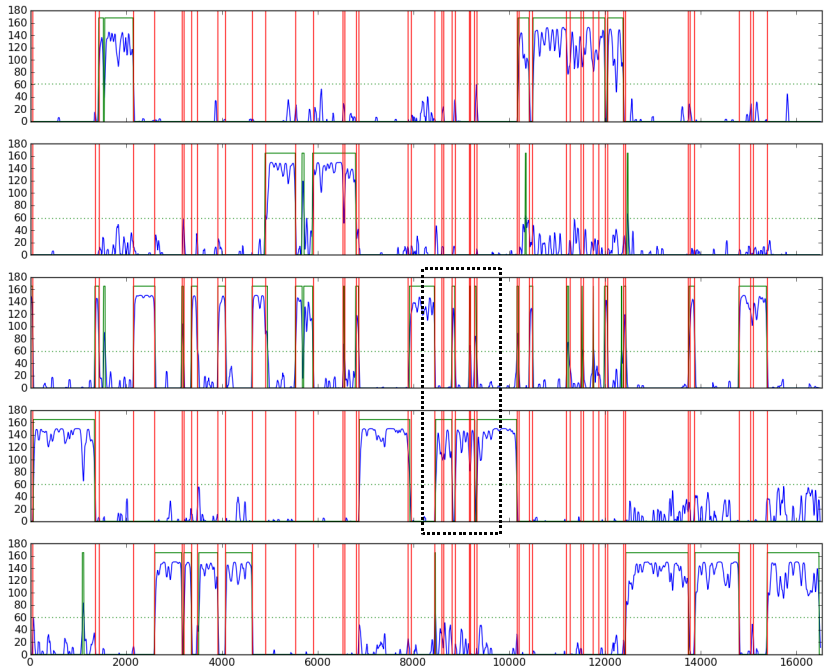

Figure 6: Convex NMF output on the audio descriptors. Red vertical lines are groundtruth speaker segments (where a new segment is created every time there is a change in the set of active speakers, hence some segments correspond to overlapped speech). Dotted green lines represent decision thresholds (here 0.4 times the maximum activation value for each component), while continuous green lines are constants representing all activation-coefficients that are above the threshold. The dotted-line rectangle highlights a region where overlapped speech occurs and the NMF components of the two corresponding speakers are activated simultaneously.

7 Conclusions

In this work we have proposed a new generic structuring paradigm whereby, whatever the modality (audio or video), NMF is applied on histogram descriptors relating to a bag of features representation, to jointly discover latent patterns, representative of elementary events, and their activations in time. Second, at the algorithmic level, we have described a majorization-minimization algorithm for novel smooth and convex variants of KL-NMF. Our approach was shown to give results clearly superior to a reference HMM system on a person-oriented video structuring application with an unpenalised standard NMF. Smoothing with a suitable value of the penalty weighting parameter β was shown to improve results even more. We have also illustrated the relevance of our general approach on a speaker diarization problem, on audio data. In that case we found our convex (and smooth) variant of KL-NMF to be necessary to obtain satisfactory results, at the expense of an increased computational burden.

A first perspective of this work would be a thorough evaluation of our approach on the audio speaker diarization task. On the methodological side, perspectives concern the automatic evaluation of the "hyperparameters", i.e., β and the number of components K. These are common issues of factorizations models, that may be handled through cross-validation or user feedback, or through Bayesian integration [3]. An other perspective is the design of online matrix factorization techniques [18] to alleviate the computational burden incurred in the large scale multimedia setting.

8 Acknowledgments

This work is supported by project ANR-09-JCJC-0073-01 TANGERINE (Theory and applications of nonnegative matrix factorization). Portions of the research in this paper use the Canal9 Political Debates Database made available by A. Vinciarelli at the Idiap Research Institute, Switzerland.

References

[1] J. Assfalg, M. Bertini, C. Colombo, and A.D. Bimbo. Semantic annotation of sports videos. *Multimedia, IEEE*, 9(2):52–60, 2002.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on Computer Vision*. IEEE, 2007.

[3] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009.

[4] A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Medical Imaging*, 12(2):328–333, 1993.

[5] C. H. Q. Ding, Tao Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45 – 55, 2010.

[6] Julian Eggert and Edgar Körner. Sparse coding and NMF. In *Proc. IEEE International Joint Conference on Neural Networks*, pages 2529–2533, 2004.

[7] Elie El Khoury, Christine Senac, and Philippe Joly. Face-and-clothing based people clustering in video content. In *International conference on Multimedia information retrieval*, page 295, Philadelphia, Pennsylvania, USA, 2010.

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

[9] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05)*, pages 601–602, New York, NY, USA, 2005. ACM.

[10] D. Guillamet, B. Schiele, and J. Vitria. Analyzing non-negative matrix factorization for image classification. In *Proc. 16th International Pattern Recognition Conference (ICPR)*, volume 2, pages 116–119, 2002.

[11] Thomas Hofman. Probabilistic latent semantic indexing. In *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999.

[12] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30 – 37, 2004.

[13] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[14] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

[15] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[16] M. Levy, M. Sandler, and M. Casey. Extraction of high-level musical structure from audio data and its application to thumbnail generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5. IEEE, 2006.

[17] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized parts-based representations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 207–212, Hawaii, USA, 2001.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:10–60, 2010.

[19] S. Meignier, J.F. Bonastre, and S. Igounet. E-HMM approach for learning and adapting sound models for speaker indexing. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, 2001.

[20] NIST. The NIST Rich Transcription 2009 (RT'09) evaluation, 2009.

[21] D. A Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 5, 2005.

[22] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?": Learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1145–1152, Miami, 2009.

[23] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS)*, St-Malo, France, apr 2009.

[24] F Vallet, S Essid, J Carrive, and G Richard. Robust Visual Features for the Multimodal Identification of Unregistered Speakers in TV Talk-shows. In *IEEE International Conference on Image Processing (ICIP)*, Hong Kong, 2010.

[25] F. Vallet, S. Essid, J. Carrive, and G. Richard. *TV Content Analysis: Techniques and Applications*, chapter High-level TV talk show structuring centered on speakers' interventions. CRC Press, Taylor Francis LLC, 2011. To appear.

[26] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[27] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *IEEE International Workshop on Social Signal Processing*, Amsterdam, 2009. Ieee.

[28] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages I–511–I–518, 2001.

[29] T. Virtanen, A. T. Cemgil, and S. Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP'08)*, pages 1825–1828, Las Vegas, Nevada, USA, Apr. 2008.

[30] M.M. Yeung and B. Liu. Efficient matching and clustering of video shots. In *IEEE International Conference on Image Processing (ICIP)*, page 338, Hong Kong, 1995.

[31] M. Zelenák and J. Hernando. On the Improvement of Speaker Diarization by Detecting Overlapped Speech. In *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, 2010.