Smooth Nonnegative Matrix Factorization for Unsupervised Audiovisual Document Structuring

Slim Essid, *Member, IEEE*, and Cédric Févotte, *Member, IEEE*

***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 by 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 two person-oriented video structuring tasks (one using the visual modality and the second the audio). This is done 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.**

***Index Terms*—Content structuring, Unsupervised classification, Machine learning, Videos, Indexing, Bag of features, Matrix factorization.**

I. Introduction

Automatic audiovisual document *structuring* represents a key technological component as part of the global effort to set up efficient multimedia and video indexing tools. Though there seems to be no consensual definition of this process, it is widely accepted that it is one of extracting a temporal organization of an audiovisual document, by organizing it into different sections, or *structural units*, each conveying a homogeneous (audio/video) type of content (possibly highlighting content repetitions). The definition of a "structural unit" highly depends both on the particular type of content that is processed and the application considered, for which a human-generated groundtruth is generally available for a set of manually annotated documents. Then, the structuring problem comes down to automatically recreating the documents temporal-organization groundtruth (obviously in view of automatically structuring new documents that have not been manually annotated). As such, *shot boundary detection* [1] or *scene segmentation*, also referred to as *sequence* [2], *story unit* [3] or *logical unit* [4] segmentation, etc., can be considered as instances of video structuring problems. Other works consider more specific structuring tasks and rely on expert techniques specifically tailored for the particular structuring scheme that is envisaged. 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) [5], [6].

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 [7]. In the audio domain they are merely abstract homogeneous content segments (hopefully belonging to different sound classes such as music, silence, speakers, etc.). These segments are generally found by a variant of the Bayesian Information Criterion technique [8].

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 (1.5s in this paper experiments) and adjacent frames typically overlap in 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) [9], or more generally nonnegative matrix factorization (NMF) with the Kullback-Leibler (KL) divergence [10], 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

Copyright (c) 2010 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

S. Essid is with Institut Telecom - Telecom ParisTech, CNRS LTCI, 37 rue Dareau - 75014 Paris, France

C. Févotte is with CNRS LTCI, Telecom ParisTech, 37 rue Dareau - 75014 Paris, France.

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. [11] 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 [12], [13]. In our setting, NMF is instead used at the classification step, after the bag of words transformation. The closest to our work is probably [14] 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 II, and expose our new NMF algorithms in Section III. We then present two distinct instantiations of our general approach on two different applications that experimentally validate its effectiveness for any particular structuring scheme, before we suggest some conclusions.

II. 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 two applications 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 to video structuring is depicted in Figure 1. Once the (audio or visual) word-histogram descriptors have been extracted, they are processed by an NMF algorithm (as explained in detail in the next section). The NMF algorithm represents the descriptors as the activations of particular basis vectors (to be associated in this application with target structural units) at every time instant. By thresholding the activations, the temporal structure of the document is deduced.

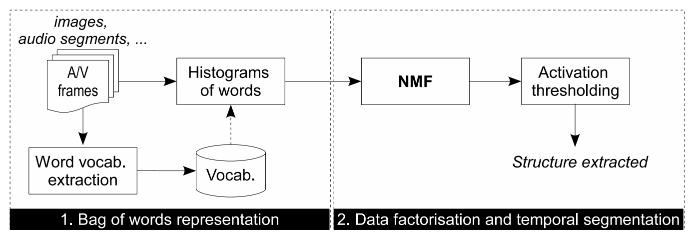


Fig. 1. General approach overview.

III. Smooth NMF for histogram sequences processing

A. Motivation

Given histogram data V with coefficients v_{fn} representing the contribution of "word" 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 "source" (e.g., 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 sources 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 [15], *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 source detection criterion on the amplitudes of the coefficients of H, using appropriate thresholding.

B. Specifications

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. [10]. More precisely, the generalized KL divergence is a minus log-likelihood in disguise for the Poisson noise model such that $v_{fn} \sim \mathcal{P}(\sum_k w_{fk}h_{kn})$, a common noise model of count data [16].

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 [17], but we here resort to the more standard smoothness measure (4) for which we will derive an original algorithm in Section IV-A.

C. Forming the objective function

In this section we present the general objective function to be minimized over W and H and show the necessity of mininization subject to fixed-norm constraints on the columns of W to prevent from degenerate solutions. Assembling the previous specifications, described in Section III-B, we are left with the following minimization problem:

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

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

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

As it appears, and following [18], [19], the minimization (6) is equivalent to the minimization of the following scale-invariant objective function:

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

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).

IV. Majorization-minimization for smooth KL-NMF

We describe a novel 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 IV-A and IV-B describe the updates of H and W, respectively. A convex-NMF variant will also later be exposed in Section VI.

A. Update of H given W

1) Unpenalized case ($\beta = 0$): In the unpenalized case and given W we are left with

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

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 [20], 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 [21], [22], [23], 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 $h_{kn} = \psi_{kn}/\lambda_k$

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

$$\begin{aligned} &L(h_n; h_{n-1}, h_{n+1}) \\ &= \frac{1}{2} \sum_k \lambda_k^2 \left[(h_{k(n+1)} - h_{kn})^2 + (h_{kn} - h_{k(n-1)})^2\right] \quad (11) \\ &= \sum_k \lambda_k^2 \left[h_{kn}^2 - (h_{k(n+1)} + h_{k(n-1)})h_{kn})\right] + cst \end{aligned}$$

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

$$G_\beta(h_n|\tilde{h}_n) = G(h_n|\tilde{h}_n) + \beta\, L(h_n; h_{n-1}, h_{n+1}). \tag{12}$$

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

$$h_{kn} = \frac{\sqrt{b_{kn}^2 + 4a_{kn}\psi_{kn}} - b_{kn}}{2a_{kn}}, \tag{13}$$

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 (13). $\tilde{V}$ is then updated with the new value $\tilde{H} = H$, and the algorithm proceeds to next iteration.

B. Update of W given H

1) Unpenalized case ($\beta = 0$): In the unpenalized case and given H, we are left with

$$\min_W C(W) = D_{KL}(V|WH) \quad s.t \quad W \geq 0. \tag{14}$$

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

$$G(W|\tilde{W}) = \sum_{fk} -\phi_{fk} \log w_{fk} + \sigma_k w_{fk} + cst \tag{15}$$

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 $w_{fk} = \phi_{fk}/\sigma_k$.

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

$$\min_W \bar{C}(W) = D_{KL}(V|WH) + \frac{\beta}{2} \sum_k s_k \lambda_k^2 \qquad s.t \quad W \geq 0 \tag{16}$$

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

$$G_\beta(W|\tilde{W}) = G(W|\tilde{W}) + \frac{\beta}{2} \sum_k s_k \lambda_k^2 \tag{17}$$

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

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

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

V. Application 1: Video structuring based on persons appearing on-screen

For a variety of TV shows, a structuring scheme centered on show-participants' occurrences is particularly meaningful and useful [6]. Thus, as a first instantiation of our generic video-structuring scheme previously presented, we consider the task of automatically categorizing each frame of talk-show videos into *"multiple participants"*, *"full group"* and *"personal shot"*, differentiating for the latter category the occurrences of each participant (hence the number of target categories is the number of persons appearing in a video plus two). This is further explained in the following.

A. Structuring-task statement and video corpus

We exploit the *Canal9 political debates* database for our application [24]. 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".

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

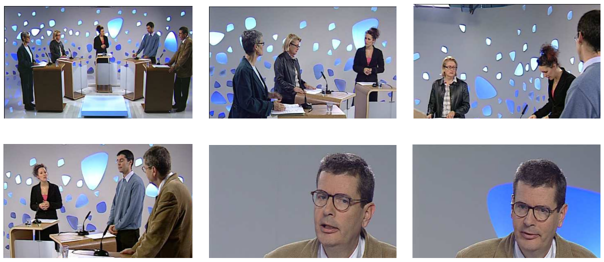


Fig. 2. Canal9 annotated shot-types. First shot (upper-left image) is labeled *"full group"*, next shots: *"multiple participants"*, and last 2 shots are *"personal shots"* labeled with the identity of the onscreen person.

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.

Our goal is 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.

To this end we instantiate our NMF-based structuring scheme as shown in Figure 3. As previously explained, this structuring scheme is completely generic and only the vocabulary creation module needs to be adapted to the particular structuring task considered, as explained hereafter.

B. *Visual vocabulary creation*

The visual vocabulary is created in such a way to be efficient for onscreen-person spotting. The latter has been considered in a number of studies [25], [26]. 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 [27], [28] 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 left corner of Figure 3. These regions are spotted as follows. First a Viola & Jones face detector [29] 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 [26]. 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]

It is worth mentioning that though color histograms seem to be natural descriptors of the clothing regions [26], [30], 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 [28]. 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[2] 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.

C. *Experimental evaluation*

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,

[1] We will see in Section V-C 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.

[2] The counts correspond to the occurrences of the vocabulary elements in the current frame, as classically done in any bag-of-words approach.

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

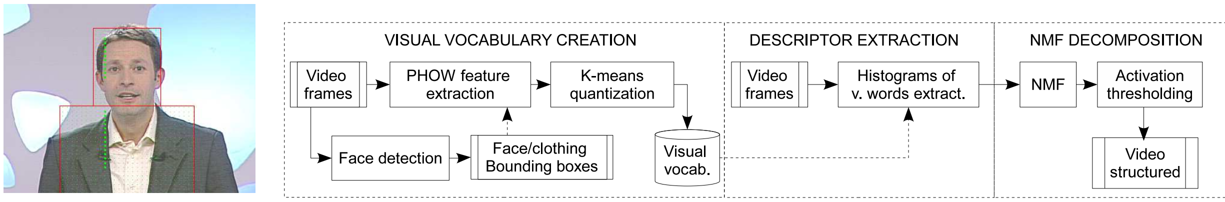


Fig. 3. Visual vocabulary creation for the first instantiation of our generic structuring scheme. Left panel: PHOW feature extraction on a visual frame during vocabulary creation phase (cf. Section V-B).

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.

1) Reference system and evaluation procedure: To the best of our knowledge, there has not been any previous works addressing the specific structuring task we are handling in this Section, hence there is no existing system to which we could compare ours. Therefore, in order to assess the performance of our proposal, we have implemented a competitive reference system that uses ergodic HMMs [31] to model the same sequence of visual-word histograms exploited by our NMF system. We keep the features fixed for both systems to make a fair comparison between the components that are responsible for the extraction of the temporal structure and better demonstrate the effectiveness of the NMF component, which is the key contribution of this paper. Note that HMMs have been successfully used in a number of previous related works, though for different applications, see for instance [32].

The HMMs we use 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. Each state is expected to represent a different structure category, hence $N_{sp}+2$ 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”. HMMs are trained using the Baum-Welch learning algorithm. The initialisation of the model parameters is done in a “standard” manner, using uniform initial and transition probabilities, and empirical means and covariances for the emission probabilities after K-means clustering [31].

Note that 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 [25]), 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 [33].

Scoring is performed following NIST[4] speaker diarization evaluation procedure[5] [34] 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.

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 4. 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 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

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

[5]We actually use the NIST scoring scripts.

Fig. 4. 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.

elements of the background as can be seen in Figure 3. 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 4). The thresholds are chosen, heuristically (by trial and error, 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 4. 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 4), 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 5 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).

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 6.

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) (best scored-video), 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.

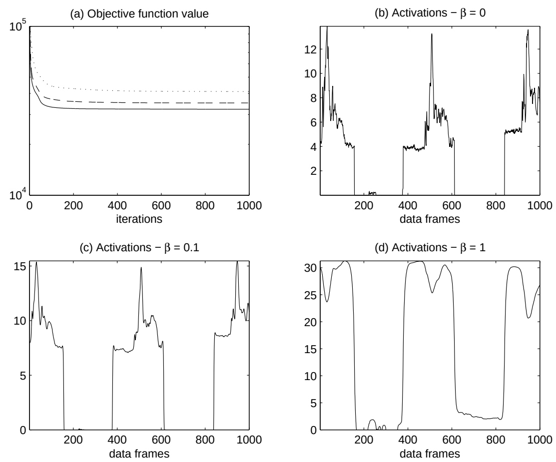


Fig. 5. Smooth KL-NMF results on video 06-10-04 (visual stream); $F = 128$, $N = 15001$ and $K = 6$. (a) Cost functions for $\beta = 0$ (solid line), $\beta = 0.1$ (dashed), $\beta = 1$ (dotted). (b-d) First 1000 coefficients of $\underline{h}_1$ obtained with the three values of β. One thousand iterations of the unpenalised and penalized algorithms take respectively 349 and 362 seconds with a MATLAB implementation on a 2.8 GHz Quad-Core Mac with 8 GB RAM.

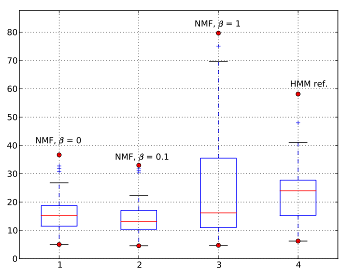


Fig. 6. Overall visual structuring error in % and box plots of the per-show visual structuring errors in %. Whiskers extend to the most extreme scores within 1.5 times the inner-quartile range.

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.

VI. Application 2: Speaker diarization

We propose a second instantiation of our generic structuring scheme on a *speaker diarization* problem using the same dataset that was used in the first application, now focusing on the audio modality (and not considering the image modality). We here merely aim to make a proof of concept (on a single show). Our goal is to emphasize that our approach can be truly applied to various tasks, and to show its potential for complex problems such as speaker diarization. For this problem, it exhibits its capacity to cope with *overlapped speech* segments, which is an issue that remains critical for researchers in this field [35].

A. 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 [36]). 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 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 [37], where we use HMMs only to build the audio descriptors and leave the speaker modeling and decision taking tasks to the NMF algorithm.

Figure VI 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 ($Q = 80$ in the experiment presented hereafter), 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).

B. NMF decomposition for speaker diarization

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 [12], 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

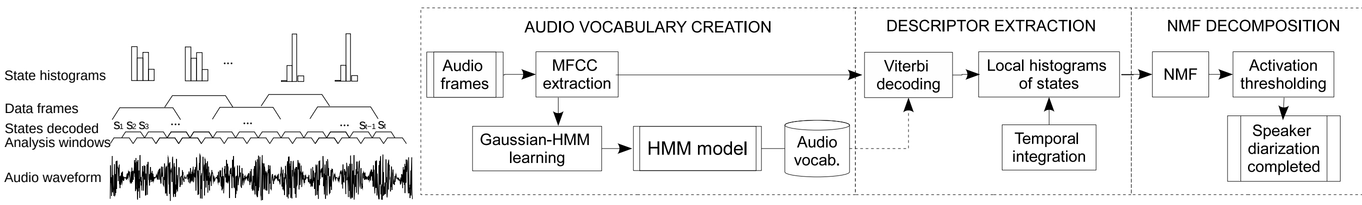


Fig. 7. Audio vocabulary creation for the second instantiation of our generic structuring scheme. Left panel: extraction of the state histogram descriptor on an audio signal (cf. Section VI-A).

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. [11], 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.* [11] consider convex-NMF with the Euclidean distance, but we obtained similar findings with the KL divergence. It is possible to combine the results of [23], which reports MM updates for convex-NMF with the KL divergence, with the results of Section IV-B to produce a MM algorithm for smooth & convex KL-NMF. Given $W = VL$, the update of H given by Eq. (13) is unchanged.

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

$$l_{mk} = \frac{\sqrt{b_{mk}^2 + 4a_{mk}\phi_{mk}} - b_{mk}}{2a_{mk}}, \tag{19}$$

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.

C. Experimental proof of concept

We validate our NMF-based approach by comparing it to a state-of-the-art speaker diarization system, namely the *LIUM SPKDIARIZATION* system [38]. The NIST speaker diarization error achieved by our proposal is 7.38% while the LIUM system error is 14.16%. Thus, our diarization scheme appears to be quite promising as it performs much better than a state-of-the-art system on the development video.

Figure 8 depicts the activations found by our convex NMF algorithm, with $\beta = 0.5$, applied to the audio-word histograms of our development video. One of the interesting features of our approach is 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

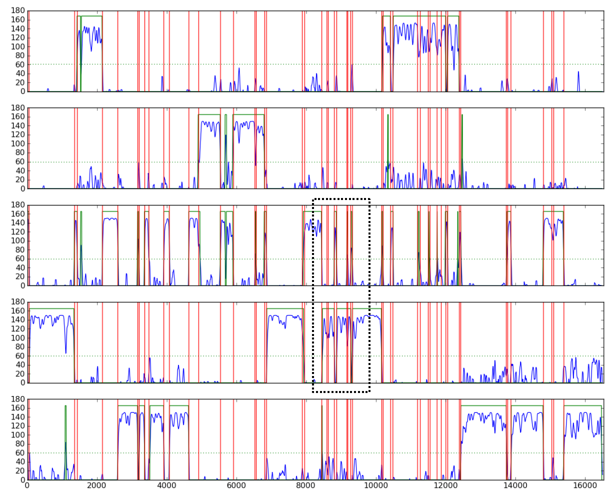


Fig. 8. Convex NMF output on the audio descriptors. Red vertical lines are groundtruth speaker segments (where a new segment is created every time there is a change in the set of active speakers, hence some segments correspond to overlapped speech). Dotted green lines represent decision thresholds (here 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.

decompose a same speaker on two different components and to represent others with the same component.

VII. 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 proposed 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 unpenalized 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, as a second instantiation of our general approach to structuring. 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 more thorough evaluation of our approach on the audio speaker diarization task. Secondly, there is interesting potential in combining the two previous applications (treated separately in this paper) into a single multi-modal system. Indeed, these visual and audio structures produce 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. There is even more potential for joint audio-visual approaches, since there is obviously a strong correlation between the audio and visual NMF activation coefficients (used to deduce a video structure). These correlations can be easily exploited to stabilize the audio activations on segments where a person is jointly speaking and appearing onscreen.

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 [39]. An other perspective is the design of online matrix factorization techniques [40] to alleviate the computational burden incurred in the large scale multimedia setting.

VIII. 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] A. Smeaton, P. Over, and A. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[2] P. Aigrain, P. Joly, and V. Longueville, "Intelligent multimedia information retrieval," M. T. Maybury, Ed. Cambridge, MA, USA: MIT Press, 1997, ch. Medium knowledge-based macro-segmentation of video into sequences, pp. 159–173. [Online]. Available: http://dl.acm.org/citation.cfm?id=266260.266272

[3] M. Yeung and B. Yeo, "Time-constrained clustering for segmentation of video into story units," in *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, vol. 3. IEEE, 1996, pp. 375–380.

[4] J. Vendrig and M. Worring, "Systematic evaluation of logical story unit segmentation," *Multimedia, IEEE Transactions on*, vol. 4, no. 4, pp. 492–499, 2002.

[5] J. Assfalg, M. Bertini, C. Colombo, and A. Bimbo, "Semantic annotation of sports videos," *Multimedia, IEEE*, vol. 9, no. 2, pp. 52–60, 2002. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=998060

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

[7] M. Yeung and B. Liu, "Efficient matching and clustering of video shots," in *IEEE International Conference on Image Processing (ICIP)*, Hong Kong, 1995, p. 338. [Online]. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/ICIP.1995.529715

[8] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 5, 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1416463

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999. [Online]. Available: http://www.cs.brown.edu/ th/papers/Hofmann-SIGIR99.pdf

[10] 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)*. New York, NY, USA: ACM, 2005, pp. 601–602.

[11] C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45 – 55, 2010.

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

[13] 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)*, Hawaii, USA, 2001, pp. 207–212.

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

[15] C. M. Bishop, *Pattern Recognition And Machine Learning*. Springer, 2008, iSBN-13: 978-0387310732.

[16] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, ser. Econometric Society Monographs. Cambridge University Press, 1998.

[17] 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)*, Las Vegas, Nevada, USA, Apr. 2008, pp. 1825–1828.

[18] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. IEEE International Joint Conference on Neural Networks*, 2004, pp. 2529–2533.

[19] 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. [Online]. Available: http://perso.telecom-paristech.fr/ fevotte/Proceedings/icassp11c.pdf

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

[21] A. R. De Pierro, "On the relation between the ISRA and the EM algorithm for positron emission tomography," *IEEE Trans. Medical Imaging*, vol. 12, no. 2, pp. 328–333, 1993.

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

[23] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011. [Online]. Available: http://arxiv.org/pdf/1010.1763v3

[24] 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: Ieee, 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5349466

[25] 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*, Miami, 2009, pp. 1145–1152. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5206513

[26] E. El Khoury, C. Senac, and P. Joly, "Face-and-clothing based people clustering in video content," in *International conference on Multimedia information retrieval*, Philadelphia, Pennsylvania, USA, 2010, p. 295. [Online]. Available: http://dblp.uni-trier.de/rec/bibtex/conf/mir/KhourySJ10

[27] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007. [Online]. Available: http://www.computer.org/portal/web/csdl/doi/10.1109/ICCV.2007.4409066

[28] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008. [Online]. Available: http://www.vlfeat.org/

[29] 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, pp. I–511–I–518, 2001. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=990517

[30] 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.

[31] L. R. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[32] J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using hmms," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 825–828.

[33] 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. [Online]. Available: http://www.tsi.enst.fr/ fevotte/Proceedings/spars09.pdf

[34] NIST, "The NIST Rich Transcription 2009 (RT'09) evaluation," 2009. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-planv2. pdf

[35] 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. [Online]. Available: http://lorien.die.upm.es/~lapiz/rtth/JORNADAS/VI/pdfs/0032.pdf

[36] S. Meignier, J. 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.

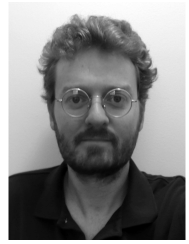
[37] M. Levy, M. Sandler, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5. IEEE, 2006.

[38] S. Meignier and T. Merlin, "LIUM SPKDIARIZATION: An open source toolkit for diarization," in *CMU SPUD Workshop*, Texas, USA, 2010. [Online]. Available: http://lium3.univ-lemans.fr/diarization/doku.php/welcome

[39] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, no. Article ID 785152, p. 17 pages, 2009, doi:10.1155/2009/785152.

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

Slim Essid is an Associate Professor at the Department of Image and Signal Processing (TSI) of Telecom ParisTech with the Audio, Acoustics and Waves group. He received the State Engineering degree from the École Nationale d'Ingénieurs de Tunis, Tunisia, in 2001, the M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 2002, and the Ph.D. degree from the Université Pierre et Marie Curie, in 2005, after completing a thesis on automatic audio classification. He has published over 50 peer-reviewed conference and journal papers with more than 50 distinct co-authors, and has served as a program committee member or as a reviewer for various audio and multimedia conferences and journals. His research interests are in machine learning for multimodal signal processing with applications in music and multimedia content analysis, and human activity/behavior analysis. He has been actively involved in various French and European funded research projects among which are Quaero, Infom@gic, NoEs Kspace and 3DLife, and collaborative projects REVERIE and VERVE.



Cédric Févotte obtained the State Engineering degree and the MSc degree in Control and Computer Science from École Centrale de Nantes (France) in 2000, and then the PhD degree in 2003. As a PhD student he was with the Signal Processing Group at Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN) where he worked on time-frequency approaches to blind source separation. From 2003 to 2006 he was a research associate with the Signal Processing Laboratory at University of Cambridge (Engineering Dept) where he worked on Bayesian approaches to sparse component analysis with applications to audio source separation. He was then a research engineer with the start-up company Mist-Technologies (now Audionamix) in Paris, designing mono/stereo to 5.1 surround sound upmix solutions. In Mar. 2007, he joined Télécom ParisTech, first as a research associate and then as a CNRS tenured research scientist in Nov. 2007. His research interests generally concern statistical signal processing and unsupervised machine learning and in particular applications to blind source separation and audio signal processing. He is the scientific leader of project TANGERINE (Theory and applications of nonnegative matrix factorization) funded by the French research funding agency ANR and a member of the IEEE "Machine learning for signal processing" technical committee.