TRACKING ATOMS WITH PARTICLES FOR AUDIO-VISUAL SOURCE LOCALIZATION

Gianluca Monaci, Pierre Vandergheynst[*]

Ecole Polytechnique Fédérale de Lausanne
Signal Processing Institute (ITS-EPFL)
CH-1015 Lausanne, Switzerland

Emilio Maggio, Andrea Cavallaro[†]

Queen Mary, University of London
Multimedia and Vision Laboratory
Mile End Road, London E1 4NS, UK

ABSTRACT

We present a general framework and an efficient algorithm for tracking relevant video structures. The structures to be tracked are implicitly defined by a Matching Pursuit procedure that extracts and ranks the most important image contours. Based on the ranking, the contours are automatically selected to initialize a Particle Filtering tracker. The proposed algorithm deals with salient video entities whose behavior has an intuitive meaning, related to the physics of the signal. Moreover, as the interactions between such structures are easily defined, the inference of higher level signal configurations can be made intuitive. The proposed algorithm improves the performance of existing video structures trackers, while reducing the computational complexity. The algorithm is demonstrated on audio-visual source localization.

Index Terms— Video signal processing, tracking, feature extraction, audio-visual processing.

1. INTRODUCTION

Object tracking is usually performed using appropriate description of the appearance of a target, either at a global or local level. Examples of global descriptions are simple templates [1], color histograms [2], or active appearance models [3]. Examples of local analysis are the methods developed to independently track and match feature points. The KLT tracker [4] first detects stable corners and then describes their appearance with an affine invariant template, computed on a small region around the corner. The points detected at subsequent frames are matched based on the appearance. More advanced feature point detectors account for rotation, scale changes of the underlying object structures [5]. All these methods are designed from a tracking-centric point of view : (i) stable structures are used to facilitate tracking, and (ii) the representation is designed to reduce ambiguity between feature points. The interpretation of the information obtained after tracking in the context of the considered signal is postponed to a subsequent analysis stage. But are stable structures also relevant from a signal representation point of view?

We argue that a signal-centric (as opposed to a tracking-centric) representation can extend the application of a feature tracking system by fusing analysis and tracking in a single general framework. The ability of tracking relevant structures of moving images would provide spatio-temporal information that is intrinsically meaningful for the representation of the video signal. Considering natural image sequences as composed of successive 2D projections of 3D objects describing smooth trajectories through time, one can assume that sequences are well modeled by smooth transformations of a reference frame. In general, a large variety of geometric structures can be found in a video sequence. A signal representation capable of exploiting video structural properties while keeping generic and flexible enough should thus be used. Such properties are introduced into the video feature extraction process, considering spatio-temporal video approximations using redundant dictionaries of geometric primitives called *atoms*. Local deformations are then propagated over time by updating the parameter field of the atoms to approximate the sequence of frames. In this framework, relevant video features are time-evolving oriented edges describing the geometric structures of a scene and their temporal evolution. An algorithm that aims at representing video sequences as a sum of relevant video structures for coding purposes was proposed in [6]. This method decomposes using Matching Pursuit (MP) a reference frame as a sparse sum of atoms taken from a redundant dictionary [7]. These structures are then tracked through time, decomposing the subsequent frames with a modified MP algorithm that uses *a priori* information inherited from previous frames. Although effective for audio-visual source localization and separation [8, 9], this video MP algorithm is formally and computationally very complex.

In this paper, we formalize the atom tracking problem to enable a more intuitive interpretation of the decomposition results and we reduce the computational complexity of the atom tracking scheme. The tracker is automatically initialized by representing the first frame of a sequence as a combination of edge-like functions. These functions are retrieved from a redundant dictionary of atoms using MP. In contrast to classical tracking algorithms, the structures to be tracked are implicitly defined by MP that picks the most relevant image contours. These visual features are then tracked with a Particle Filter (PF) [10]. The proposed scheme is demonstrated on audio-visual source localization.

The paper is organized as follows. Section 2 presents the video representation framework based on MP, and Sec. 3 the tracking algorithm based on PF. In Sec. 4 we comment the experimental results on audio-visual source localization based on edge tracking. Finally, in Sec. 5 achievements and future research directions are discussed.

2. GEOMETRIC VIDEO REPRESENTATION

Each video frame is decomposed into a low-pass part, that takes into account the smooth image components, and a high-pass part, where most of the energy of edge discontinuities lays. Assuming that this high-pass image $I(x, y)$ can be approximated with a linear combination of functions $G_{\mathbf{x}}(x, y)$ retrieved from a redundant dictionary

[*]The authors acknowledge the support of the Swiss National Science Foundation through the IM.2 National Center of Competence for Research.

[†]The authors acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC), under grant EP/D033772/1.

$\mathcal{D}_\mathcal{V}$ of 2D atoms, we can write :

$$I(x,y) \approx \sum_{\mathbf{x}[n]\in\Omega} c_{\mathbf{x}[n]} G_{\mathbf{x}[n]}(x,y)\,, \tag{1}$$

where n is the summation index, $c_\mathbf{x}$ corresponds to the coefficient for every atom $G_\mathbf{x}(x,y)$ and Ω is the subset of selected atom indexes from $\mathcal{D}_\mathcal{V}$.

The codebook $\mathcal{D}_\mathcal{V}$ is built by applying a set of geometric transformations to a mother function $G(x,y)$, in order to generate an overcomplete set of primitives spanning the input image space. The considered transformations are anisotropic scaling s_x and s_y, translations t_x and t_y, and rotation θ. The generating function G should well represent edges; thus we use an edge-detector atom that is shaped as a Gaussian along one axis and the first derivative of a Gaussian along the perpendicular one. To decompose $I(x,y)$ over the codebook $\mathcal{D}_\mathcal{V}$ we use MP, that iteratively retrieves the element of the dictionary that best matches the signal.

We consider an approach where 2D primitives $G_\mathbf{x}(x,y)$ of the form of (1), obtained in the expansion of a reference frame $I_1(x,y)$, are tracked from frame to frame (The reference frame is the first frame of the sequence). The first step of the MP algorithm decomposes I_1 as

$$I_1 = \langle I_1, G_{\mathbf{x}[0]}\rangle G_{\mathbf{x}[0]} + R^1 I_1\,, \tag{2}$$

where $R^1 I_1$ is the residual component after approximating I_1 in the subspace described by $G_{\mathbf{x}[0]}$. $G_{\mathbf{x}[0]}$ is chosen such that the projection $|\langle I_1, G_{\mathbf{x}[0]}\rangle|$ is maximal. This procedure is recursively applied, and after N iterations we approximate I_1 as

$$I_1 = \sum_{n=0}^{N-1} c_{\mathbf{x}[n]} G_{\mathbf{x}[n]} + R^N I_1\,, \tag{3}$$

where $c_{\mathbf{x}[n]} = \langle R^n I_1, G_{\mathbf{x}[n]}\rangle$, $R^0 = I_1$ and $R^n I_1$ is the residual after n iterations. In this way the reference frame I_1 is decomposed into N atoms $G_{\mathbf{x}[n]}$ that are tracked through time.

3. TRACKING VIDEO ATOMS USING PARTICLES

Tracking is performed using Particle Filter (PF), a parametric approach that solves non-linear and non-Gaussian state estimation problems [10] and can deal with multi-modal *pdf*s.

Let the reference image be represented with N atoms. If N is relatively small, then the atoms can be tracked *independently*. The assumption of independence is motivated by the fact that we are interested only in the main video structures (i.e., the first functions of the MP decomposition). If few atoms are considered, then their interactions are likely to be weak. These interactions can be estimated by computing the scalar products between atoms: strong interactions correspond to large scalar products (since atoms have unit norm the maximum scalar product is 1), whereas weak interactions correspond to small scalar products (i.e., close to 0). Figure 1 [Left] shows the sum of the scalar products between atoms on the first frame of a test clip as a function of N. The total scalar product slowly increases with the number of functions. In our experiments we will consider the first $N = 30$ atoms selected by MP : as a first approximation, it seems reasonable to track them independently. However, neighboring functions can influence each other ([11]) and future developments of this work will account for interactions between atoms.

Each atom $G_{\mathbf{x}[n]}$ is fully characterized by the set of parameters $\mathbf{x}[n]$ in a 5D state space spanned by position, scale and rotation parameters that describe its shape. PF solves the tracking problem based on the state equation, $\mathbf{x}_t[n] = \mathbf{f}_t(\mathbf{x}_{t-1}[n], \mathbf{v}_t)$, and on

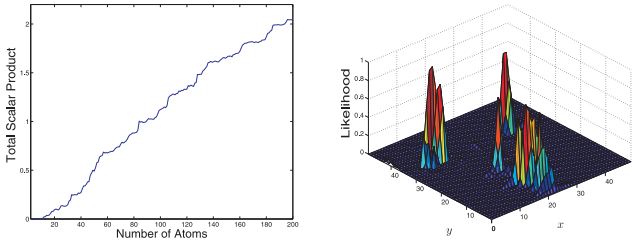

Fig. 1. Sum of scalar products between the atoms representing one frame plotted as a function of the number of atoms [Left], and likelihood of a candidate atom computed on a region extracted from one of the analyzed clips [Right].

the measurement equation, $\mathbf{z}_t[n] = \mathbf{h}_t(\mathbf{x}_t[n], \mathbf{n}_t)$, where $\mathbf{f}_t$ and $\mathbf{h}_t$ are non-linear and time-varying functions. The state variable $\mathbf{x}_t$ describes the characteristics of target n at time t, thus it defines the n-th atom at frame t. To simplify the notation, the atom index n will be omitted, since the atoms are tracked independently. $\{\mathbf{v}_t\}_{t=1,\ldots}$ and $\{\mathbf{n}_t\}_{t=1,\ldots}$ are assumed to be i.i.d. stochastic processes. The problem consists in calculating the *pdf* $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ at each time instant t. This *pdf* can be obtained recursively in two steps, namely prediction and update. PF approximates the densities $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ with a sum of N_s Dirac functions centered in $\left\{\mathbf{x}_t^i\right\}_{i=1,\ldots,N_s}$ as

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N_s} \omega_t^i \delta\left(\mathbf{x}_t - \mathbf{x}_t^i\right)\,, \tag{4}$$

where ω_t^i are the weights associated to the particles :

$$\omega_t^i \propto \omega_{t-1}^i \frac{p(\mathbf{z}_t|\mathbf{x}_t^i)p(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i)}{q(\mathbf{x}_t^i|\mathbf{x}_{t-1}^i, \mathbf{z}_t)}\,. \tag{5}$$

The function $q(\cdot)$ is the importance density function which is often chosen to be $p(\mathbf{x}_t|\mathbf{x}_{t-1}^i)$, as it is done here. This leads to $\omega_t^i \propto \omega_{t-1}^i p(\mathbf{z}_t|\mathbf{x}_t^i)$. A re-sampling algorithm can then be applied to avoid the degeneracy problem [10]. In this case, the weights are set to $\omega_{t-1}^i = 1/N_s \;\forall\, i$, and therefore $\omega_t^i \propto p(\mathbf{z}_t|\mathbf{x}_t^i)$. The weights are thus proportional to the *likelihood* of the measurement $\mathbf{z}_t$, given the particles. Here the natural choice for the likelihood function is the projection of the candidate atom over the image, since we want to track important video structures, i.e. video atoms exhibiting high projection on the image. This is also coherent with the representational framework formulated in the previous section. The likelihood of a candidate particle is defined as the absolute value of the scalar product between the residual frame and the atom represented by the particle. In order to favor candidates with high likelihood, this quantity is filtered with a Gaussian kernel centered in the maximum likelihood value and with variance $\sigma_\mathcal{L}$, thus obtaining:

$$\mathcal{L}(\mathbf{x}_t^i[n]) = \exp\left(-\frac{(\mathcal{L}_t^M[n] - |\langle R^n I_t, G_{\mathbf{x}_t^i[n]}\rangle|)^2}{2\cdot(\sigma_\mathcal{L}\mathcal{L}_t^M[n])^2}\right)\,, \tag{6}$$

with $\mathcal{L}_t^M[n] = \max(|\langle R^n I_t, G_{\mathbf{x}_t^i[n]}\rangle|)\,, i = 1,\ldots,N_s$.

We want to highlight that the atom $G_{\mathbf{x}_t^i[n]}$ is not projected over the frame I_t but over the residual at step n of the decomposition, $R^n I_t$ (see (3)). Figure 1 [Right] shows the likelihood function of a candidate atom computed on a region extracted from a test clip.

The best state at the time t, $\hat{\mathbf{x}}_t$, is the particle $\mathbf{x}_t^i$ with largest weight, pondered by a factor that takes into account the similarity of the particle with the corresponding best state at time $t-1$:

$$\hat{\mathbf{x}}_t = \mathbf{x}_t^M \quad \text{s.t.} \quad \omega_t^M = \max(s(\mathbf{x}_t^i, \hat{\mathbf{x}}_{t-1}) \cdot \omega_t^i) \,. \tag{7}$$

The function s is a Gaussian in the 5D parameters space. The value of $s(\mathbf{x}, \mathbf{y})$ is maximum when the particles $\mathbf{x}$ and $\mathbf{y}$ coincide and it decreases exponentially as the distance between $\mathbf{x}$ and $\mathbf{y}$ in the parameters space increases.

Alternative strategies to compute the best state would be to take the particle with highest weight or to consider the Monte Carlo approximation of equation (4), i.e. the weighted sum of the particles [10]. However, we observed that unstable, noisy atom trajectories were generated considering simply the particles with largest weights, due to the multi-modality of the posterior *pdf*s (see Fig. 1 [Right]). The Monte Carlo solution produces more stable atom trajectories, but in this case there is no guarantee that the best state corresponds to an atom that matches a *real* visual structure, since several local maxima can be present in the likelihood function (Fig. 1 [Right]). The introduction of the weight $s(\mathbf{x}, \mathbf{y})$ stabilizes the atoms tracks since the algorithm tends to prefer states that are as similar as possible to the previous ones, except if relevant modifications occur. At the same time, the representation of the scene is kept coherent.

4. EXPERIMENTAL RESULTS

In this section we present the results of the atom tracking method using PF (MP-PF). We test the algorithm on sequences representing one or two persons speaking and moving in front of a camera, taken from the CUAVE database [12][1]. The video data is at 29.97 fps and at a resolution reduced from 480 × 720 to 120 × 176 pixels. We use a 5-dimensional state model for PF, composed of the target position, (x, y), the target size s_x and s_y and the orientation θ. In all experiments we use a zero-order motion model with fixed $\sigma_{t_x} = \sigma_{t_y} = 2$, $\sigma_{s_x} = \sigma_{s_y} = 0.03$ and $\sigma_\theta = 3.5$. Note that the position change is in pixels while the scale is in percentage and the orientation in degrees. The Gaussian used to filter the likelihood function has $\sigma_{\mathcal{L}} = 0.05$. The PF tracker uses 150 samples.

In the first experiment, the proposed MP-PF approach is tested on four sequences representing one person speaking and moving in front of the camera and it is compared with the video MP algorithm [6] (3D-MP). Sample frames of two clips are shown in Fig. 2. Both trackers are initialized with the same video atoms using MP as described in Sec. 2. The edges are then tracked using a video MP approach in 3D-MP, while our proposed method tracks the video structures using PF as detailed in Sec. 3. In Fig. 2 the tracking results using the two algorithms are compared. The first and third rows show the results obtained with the 3D-MP approach, while the second and forth rows show the results for the proposed MP-PF method. In the second part of the sequence (second and third frames) the subjects rapidly move towards the left. The 3D-MP tracker looses the track of two edges in the first case and of one in the second, while the MP-PF tracker does not. The same behavior has been observed in the other test sequences. While the 3D-MP algorithm easily loose the track of fast moving edges, the MP-PF approach results more robust, even if errors can be observed. In both sequences for example it happens that the yellow atom associated with the upper lip is temporarily associated with the lower lip or the chin.

In the second experiment, MP-PF is integrated in the audio-visual fusion algorithm [8] to perform a source localization task.

[1]Only the luminance component of the video clips has been considered.

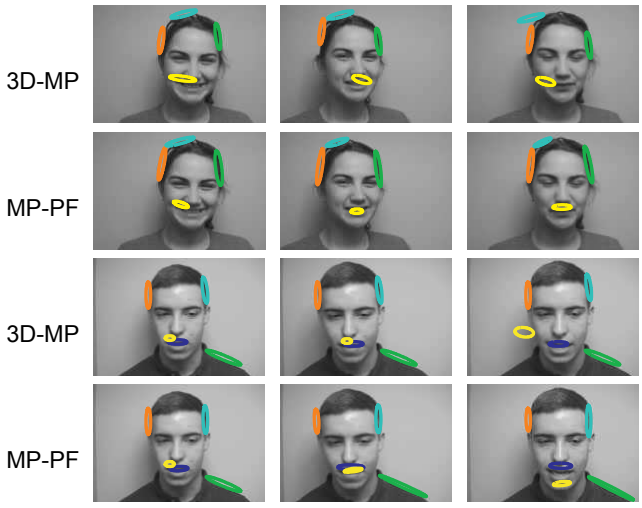

Fig. 2. Video atoms tracking. Footprints of different atoms are depicted with different colors. Results for the 3D-MP approach are on the first and third rows and those for the MP-PF method are on the second and forth rows. From the second to the third frame the subjects rapidly move towards their left : the 3D-MP tracker looses the track of some edges, while the MP-PF tracker does not.

The audio-video features that are considered here are the same used in [8, 9]. The audio signal is represented by a mono-dimensional feature that estimates the average acoustic energy. The video signal instead is represented using $M = 30$ video atoms and each atom has a feature associated describing its displacement. Peaks are extracted from audio and video features and *synchronization vectors* are built [8]. The video atoms exhibiting the highest degree of correlation with the audio are detected using a simple relevance criterion and the sound source location over the image sequence is estimated. A sliding window of 70 frames length is used to compute the synchronization vectors and to detect the video atoms that are more correlated with the audio. The observation window is then shifted by 20 samples and the procedure iterated.

We have tested the algorithm on four sequences of the CUAVE database (g19, g20, g21, g22) involving two persons reading series of digits in English. Figure 3 shows the results of the proposed approach detecting the mouth of the speaker in two sequences where two persons speak in turns in front of the camera. In white are highlighted the footprints of the atoms found to be correlated with the soundtrack. The mouths of the correct speakers are detected.

To quantify the accuracy of the method, the center of the speaker's mouth in the test sequences has been manually labeled, and the detection performances compared with those of two cross-modal source localization algorithms [8, 13]. In [13] a method is proposed to detect the mouth of the speaker founding the image zone over which the mutual information between audio and video features is maximized. As already stated, here we use the same scheme as in [8], with the difference that in [8] the 3D-MP approach is used to track the video atoms.

The active speaker's mouth is considered to be correctly detected if the position of the most correlated video atom falls within a circle of 50 pixels of diameter centered in the labeled mouth center. All methods detect correlated video structures every 20 frames and thus performance is evaluated with this same frequency. Table 1 sum-

Fig. 3. Frames from clips g19 [Top] and g21 [Bottom]. The footprints of the most correlated atoms are highlighted. The mouths of the correct speakers are detected.

marizes the results obtained for the three methods in term of percentage of test points at which the speaker's mouth is correctly detected. The use of geometric video decompositions combined with an audio-video event detector is confirmed to improve the results obtained maximizing mutual information ([13]). The proposed method has a detection performance similar to that of Monaci's algorithm, slightly improving previous results for sequence g19 but obtaining inferior performances on clip g22.

The MP-PF method improves the tracking performances of the 3D-MP tracking algorithm, as shown by the results in Fig. 2. This is indeed interesting considering that the 3D-MP algorithm, even without jointly tracking groups of structures, takes into account atoms' interactions, which was demonstrated to increase the accuracy of the 3D-MP approach [11]. We argue that a MP-PF algorithm that takes into account atoms' dependencies would correct tracking errors due to atoms' interactions (Fig. 2) and would allow to improve the audio-visual localization results, that by now are essentially equivalent to those obtained using 3D-MP (Table 1). Concerning the computational complexity, we have tested the two methods on a video sequence whose 30 principal video atoms were tracked through time. The MP-PF algorithm clearly outperforms the 3D-MP approach, resulting approximately 7 times faster.

5. DISCUSSION

We presented a new framework and an efficient algorithm to represent and track relevant video structures. The proposed method improves the 3D-MP video representation algorithm presented in [6], which is designed as a coding algorithm and poses problems from the tracking point of view. The parameters of the video atoms are in fact coarsely quantized to achieve better compression performances, introducing tracking errors. Moreover, atoms are tracked using a search window of reduced size, which limits the robustness and accuracy of the tracker. These limitations are overcome by defining the video atom tracking problem in the well grounded and understood framework of PF, which ensures robustness, flexibility and lower computational complexity than the 3D-MP algorithm.

Experiments show that the proposed tracker is more robust and accurate than the 3D-MP one, while being considerably less time consuming. The audio-visual source localization algorithm, however, does not improve accordingly. This is mainly due to the fact that while in [8] the 3D-MP algorithm takes into account atoms' interactions, the current MP-PF method does not. This in certain situations produces less stable atoms trajectories because of interferences between atoms, as shown in Fig. 2. However these results show that there is room for further improvements by designing a mechanism that accounts for the interactions between video atoms. The tracking framework developed in this paper seems to be appropriate to continue the evolution of our system.

Clip	Nock[13]	Monaci[8] (3D-MP)	Proposed (MP-PF)
g19	41	87	94
g20	93	93	93
g21	79	81	78
g22	79	87	80

Table 1. Results expressed in percentage of correct detections.

6. REFERENCES

[1] Y.-S. Yao and R. Chellappa, "Tracking a dynamic set of feature points," *IEEE Trans. Image Proc.*, vol. 4, no. 10, pp. 1382–1395, 1995.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 5, pp. 564–577, May 2003.

[3] G. Edwards, C. Taylor, and T. Cootes, "Interpreting face images using active appearance models," in *Proc. 3^{rd} Int. Conf. Automatic Face and Gesture Recogn.*, 1998, pp. 300–305.

[4] B.D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. IJCAI*, 1981, pp. 674–679.

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] Ò. Divorra Escoda, *Toward Sparse and Geometry Adapted Video Approximations*, Ph.D. thesis, EPFL, Lausanne, June 2005, [Online] Available: `http://lts2www.epfl.ch/`.

[7] S. Mallat and Z. Zhang, "Matching Pursuits with time-frequency dictionaries," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[8] G. Monaci and P. Vandergheynst, "Audiovisual gestalts," in *Proc. of CVPR Workshop POCV*, 2006.

[9] G. Monaci, Ò. Divorra Escoda, and P. Vandergheynst, "Analysis of multimodal sequences using geometric video representations," *Signal Processing*, vol. 86, no. 12, pp. 3534–3548, 2006.

[10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Proc.*, vol. 50, no. 2, pp. 174–188, 2002.

[11] Ò. Divorra Escoda and P. Vandergheynst, "A Bayesian approach to video expansions on parametric over-complete 2-D dictionaries," in *Proc. of IEEE MMSP*, 2004, pp. 490–493.

[12] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP JASP*, vol. 2002, no. 11, pp. 1189–1201, 2002.

[13] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localisation using audio-visual synchrony: an empirical study," in *Proc. Int. Conf. Image Video Retrieval (CIVR)*, 2003, pp. 488–499.