UNSUPERVISED EXTRACTION OF AUDIO-VISUAL OBJECTS

Anna Llagostera Casanovas and Pierre Vandergheynst

Signal Processing Institute (LTS2), École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

We propose a novel method to automatically detect and extract the video modality of the sound sources that are present in a scene. For this purpose, we first assess the synchrony between the moving objects captured with a video camera and the sounds recorded by a microphone. Next, video regions presenting a high coherence with the soundtrack are automatically labelled as being part of the source. This represents the starting point for an innovative video segmentation approach, whose objective is to extract the complete audio-visual object. The proposed graph-cut segmentation procedure includes an audio-visual term that links together pixels in regions with high audio-video coherence. Our approach is demonstrated on challenging sequences presenting non-stationary sound sources and distracting moving objects.

Index Terms— audio-visual processing, graph cuts

1. INTRODUCTION

After the preliminary work of in Hershey and Movellan in [1], numerous approaches performed a joint analysis of information in audio and video modalities in order to locate the sound sources in the image [2, 3]. In contrast, only the method in [4], and the works of Liu and Sato in [5, 6] attempted the extraction of the source's video part. In [4] the video signal is decomposed into basic image structures (atoms), then the sources position is estimated by clustering together atoms with high audio-visual correlation, and finally each source is reconstructed by adding the contribution of the atoms that are close to its estimated position. Thus, in [4] the particular shapes of the sources are not considered, i.e. the extracted sources have always an approximately circular shape because all atoms inside a radius are used in the source reconstruction process. In [5, 6] they overcome this limitation by using a segmentation technique based on graph cuts, which is initialized by audio-visual analysis. In [5] the source position is estimated by computing the Quadratic Mutual Information between audio and video features, and this procedure is applied to sequences composed of almost static speakers. Then, in [6] this method is generalized to non-stationary sound sources by identifying the pixel's visual trajectories whose changes in acceleration better fit the energy variations in the audio channel.

The method that we present can also be applied to non-stationary sound sources. First, regions presenting a high coherence with the audio channel are automatically assigned to the audio-visual object. Then, the remaining pixels are binary classified into object or background by using a novel audio-visual graph-cut segmentation procedure that keeps together pixels in regions presenting a high coherence with the soundtrack. Between all segmentation techniques graph cuts have shown applicability to N-dimensional problems and flexibility in the definition of the energy to minimize, for which they provide a globally optimal segmentation through a numerically robust minimization procedure. They were first introduced in [7] for monochrome N-D signals and extended to color images and videos in latter approaches [8, 9].

Let us now detail the main contributions of our approach:

1. From a video segmentation point of view, the introduction of *audio-visual priors* makes the segmentation automatic. The necessity of user interaction is the main limit of previous segmentation approaches [7, 9, 8].
2. We propose an innovative *audio-visual term* in the energy function that the graph cut algorithm minimizes. This term links together neighboring pixels presenting a high audio-visual coherence and thus probably belonging to the audio-visual object. Unlike in [5, 6], our audio-visual term does not affect regions with low coherence and thus it does not include any implicit assumption about these regions. The term in [5, 6] forces the regions presenting low correlation with the soundtrack to be part of the background. As a result, in our case the audio-visual object can be completely extracted even though some parts of it present a lower audio-visual coherence.
3. We redefine the standard *regional term* in the segmentation's energy function, which integrates knowlegde about the color distributions in foreground and background. In Sec. 3 we demonstrate the advantages of the proposed regional term over the commonly adopted term in [7, 9, 8]. Furthermore, keeping this term represents a significant advantage over the methods in [5, 6], since it ensures the cohesion between the homogeneous regions composing the audio-visual object.

The paper is structured as follows. In Sec. 2 we define the audio-visual coherence, a measure to quantify the relationship between video structures and sounds at the pixel level. Sec. 3 explains the 3D graph cut segmentation of a group of frames (GoF), which integrates the knowledge from joint audio-visual analysis. In Sec. 4 we present an automatic criteria to choose the segmentation priors according to the audio-visual coherence. Sec. 5 presents the results obtained on challenging audio-visual sequences. In Sec. 6 achievements and future research directions are discussed.

2. AUDIO-VISUAL COHERENCE

In a first stage the audio-visual diffusion process presented in [10] is used to assess the correlation between audio and video channels. This nonlinear diffusion procedure reduces the information (spatio-temporal edges) in video regions whose motion is not coherent with the soundtrack. Thus, we can easily deduce the regions in which the video signal is least diffused by simply comparing the motion (temporal edges) before and after the audio-visual diffusion process. The regions in which the motion is better preserved are, with high probability, part of the audio-visual object since their movements are correlated to the sounds in the audio channel.

This work is funded by the Swiss NFS through grant number 200021-117884 and by the EU Framework 7 FET-Open project FP7-ICT-225913-SMALL: Sparse Models, Algorithms and Learning for Large-Scale data.

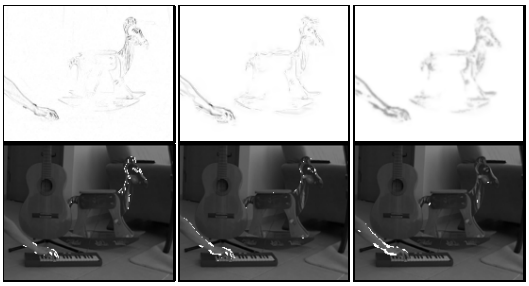


Fig. 1. White pixels in the bottom indicate the 0.5% highest values of the top features: [from left to right] original motion $\partial_t v(\mathbf{x}, 0)$, resulting motion $\partial_t v(\mathbf{x}, \tau_{stop})$ and audio-visual coherence $c(\mathbf{x})$. A hand is playing a synthesizer while a rocking horse is moving.

Let $v(\mathbf{x}, \tau)$ be the video signal v at spatio-temporal coordinates $\mathbf{x}$ and diffusion time τ. We define the *audio-visual coherence* $c(\mathbf{x}) \in [0, 1]$ at pixel location $\mathbf{x}$ as

$$c(\mathbf{x}) = \begin{cases} \frac{1}{s}\frac{\partial_t v(\mathbf{x},\tau_{stop})}{\partial_t v(\mathbf{x},0)} & \text{if} \quad \partial_t v(\mathbf{x},0) > \xi \\ \frac{1}{s}\frac{\partial_t v(\mathbf{x},\tau_{stop})}{\text{argmax}_{\mathbf{x}}\, \partial_t v(\mathbf{x},0)} & \text{else} \end{cases} \tag{1}$$

where $\partial_t v(\mathbf{x}, \tau_{stop})$ is the temporal derivative of the resulting video signal after n_{stop} iterations of the proposed nonlinear diffusion procedure ($\tau_{stop} = n_{stop}\Delta\tau$), the constant ξ makes the audio-visual coherence $c(\mathbf{x})$ close to zero in static pixels (we can fix $\xi = 10^{-1}$ for example), and the constant s makes $c(\mathbf{x})$ unitary. Thus, the higher is the audio-visual coherence $c(\mathbf{x})$ the higher is the probability for the video pixel at location $\mathbf{x}$ to be part of an audio-visual object, since its motion is well preserved through the diffusion process.

Fig. 1 shows a frame of a sequence where the video motion in the audio-visual object has approximately the same magnitude than the distracting motion (the highest values of the original motion are equally distributed between hand and horse). After the audio-visual diffusion process the motion is already more intense in the hand region [center], while the audio-visual coherence [right] is clearly dominant in the audio-visual object (the hand's silhouette is darker [top] and only a few white pixels appear over the rocking horse).

Thus, the *audio-visual coherence* represents an efficient measure of the relationship between video regions and the audio signal, with a high spatial resolution. This measure is used in Sec. 3 in the definition of the audio-visual segmentation problem and in Sec. 4 as a starting point for the proposed segmentation procedure.

3. GRAPH CUT SEGMENTATION BY EXPLOITING AUDIO-VIDEO SYNCHRONY

Our 3D segmentation approach is based on the procedure presented in [7]. Given some initial information about foreground and background locations provided by the user (seeds) they compute a globally optimal segmentation of monochrome 3D volumes using graph cuts. In this section, this procedure has been extended to color video signals by integrating joint audio-visual processing.

Let $\mathbf{z} = (z_1, \ldots, z_p, \ldots, z_P)$ be the set of pixels in the RGB color space that compose a group of frames (GoF). The segmentation process consists on assigning a binary label $l = (l_1, \ldots, l_P)$ to each pixel p: $l_p \in \{0(\text{background}), 1(\text{foreground})\}$.

First, we build a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ corresponding to a GoF following the procedure in [7]. The set of vertices $\mathcal{V}$ is composed of the P pixels in the GoF plus the foreground F and background B terminals. The set of edges $\mathcal{E}$ is composed by edges connecting neighboring pixels $\{p, q\} \in \mathcal{N}$ (n-links) and edges connecting each pixel p to the foreground and background terminals $\{p, F\}$ and $\{p, B\}$ (t-links). The neighborhood $\mathcal{N}$ of each pixel is composed of six pixels, four spatial neighbors and two temporal neighbors as in [7]. Then, the graph cut algorithm solves our segmentation problem by minimizing the following energy defined on the graph:

$$\begin{aligned} J(l) &= \lambda_R R(l) + V(l) + \lambda_C C(l) \\ &= \lambda_R \sum_{p \in \mathcal{P}^j} R_p(l_p) + \sum_{\{p,q\} \in \mathcal{N}} (V_{p,q} + \lambda_C C_{p,q})\, [l_p \neq l_q] \ , \end{aligned} \tag{2}$$

where $[\Phi]$ denotes the indicator function taking values 0, 1 for a predicate Φ. The regional term $R(l)$ evaluates how the color z_p corresponding to each pixel p with label l_p fits into the background and foreground models, the boundary term $V(l)$ assesses the similarity of each pixel with its neighborhood, and the audio-visual term $C(l)$ links together neighboring pixels belonging to a region with high audio-visual coherence. Then, the coefficients λ_R and λ_C define the relative importance of the regional term and the audiovisual term with respect to the boundary term. In all experiments this parameters have been fixed to $\lambda_R = 0.05$, a value within the range defined by [8] and [9], and $\lambda_C = 0.6$ so that the extracted region respects the strong edges in the image.

As explained before, Liu and Sato introduced an energy term that included audio-visual knowledge to extract the speaker face region [5] or general sound sources [6]. In a first stage, the Expectation Maximization algorithm was used to cluster the audio-visual correlation values into two clusters representing the sound source and the background. Then, they proposed to replace the standard regional term $R(l)$ in equation (2) by a cost to assign a pixel to be part of the sound source, which depended on the Mahalanobis distance between the pixel and the estimated mean value of the source's correlation. Here in contrast, we propose to keep the regional term (by redefining the one in [7, 8, 9]) and then introduce an audio-visual term. Our term links together neighboring pixels in regions with high audio-visual coherence instead of linking each pixel to the foreground and background terminals. Thus, the proposed term ensures that the pixels composing the audio-visual object are kept together in the segmentation process, and it does not affect regions with low coherence (they were assumed to belong to the background in [5, 6]). Since the connections between pixels are spatio-temporal, we reinforce also the links between neighboring frames in regions where the image structures move coherently with the sounds.

The *boundary term* is defined by

$$V_{p,q} = \frac{1}{dist(p,q)} \exp\left(-\frac{\|z_p - z_q\|^2}{2\gamma_V^2}\right) , \tag{3}$$

where $\gamma_V^2 = \mathrm{E}(\|z_p - z_q\|^2)$ as in [9]. Here $\mathrm{E}(\cdot)$ denotes the expectation operator over the video signal and $dist(\cdot)$ is the Euclidean distance between neighboring pixels.

Gaussian Mixture Models (GMMs) are estimated for foreground (Λ^f) and background (Λ^b) color distributions from the available seeds, by using the Expectation Maximization algorithm: $\Lambda^m = \{u_i^m, \mu_i^m, \Sigma_i^m\}_{i=1}^{Q}$ for $m = \{b, f\}$. For each Gaussian i composing the mixture, u_i, μ_i and Σ_i denote respectively its weight, mean and covariance matrix. The number of Gaussians is fixed to $Q = 5$ as in [9]. According to these color models, the penalties for assigning the pixel p to foreground ($l_p = 1$) and background ($l_p = 0$) that compose the *regional term* are defined respectively as

$$\begin{aligned} R_p(l_p = 1) &= h(\ln \mathrm{P}(z_p|\Lambda^b)) \, , \\ R_p(l_p = 0) &= h(\ln \mathrm{P}(z_p|\Lambda^f)) \, , \end{aligned} \tag{4}$$

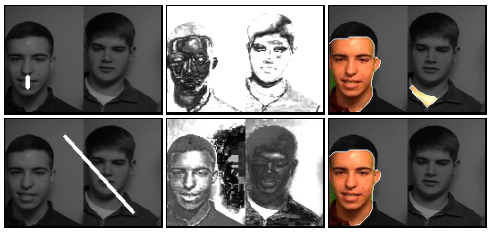


Fig. 2. Segmentation results [right] when using the regional term in previous methods [top] and our regional term [bottom] given the manually-added seeds [left] and corresponding probability maps [center] for foreground [top] and background [bottom]. No audio-visual term is used in this comparison ($\lambda_C = 0$). The foreground is shown in brighter grayscale. White regions represent the seeds [left] and a low probability [center].

where $\mathrm{P}(z_p|\Lambda^m)$ is the probability for a pixel p to belong to the foreground/background given the GMM Λ^m, and $h(\cdot)$ is a function that maps $\ln \mathrm{P}(z_p|\Lambda^m)$ from $(-\infty, 0]$ to $[0, 1]$ where "0" and "1" represent the lowest and the highest probability respectively. Thus, the weight of the edge that links any pixel p to the foreground (background) is proportional to the probability for its color z_p of belonging to the foreground (background) color model expressed by Λ^f (Λ^b). Previous methods [7, 8, 9] used the negative log-likelihoods, and thus the edge's weight was *inversely* proportional to this probability. Fig. 2 illustrates the advantages of our regional term. The probability for a pixel situated in the right person's shirt of belonging to both foreground and background is very low (in white in the central figures). According to the proposed regional term, the links between those pixels and the background and foreground terminals have a very low weight and thus they do not influence the segmentation results. However, when using the term in [7, 8, 9] the link between the pixels in the shirt and the foreground terminal is much stronger than the link to the background because the probability of belonging to the background is lower. Notice that the segmentation result contains the right person's shirt when applying the regional term in [7, 8, 9] [top], while it is not extracted in our case [right]. Thus, the regional term in previous methods enforced the segmentation algorithm to label those pixels as foreground, even though this is not clear at all according the color models. In this work, we prefer to rely on the boundary term and do not influence the segmentation when the probabilities of belonging to foreground and background are so remote.

The proposed *audio-visual term* is defined by

$$C_{p,q} = \frac{1}{dist(p,q)}\, c_p\, \exp\left(-\frac{|c_p - c_q|^2}{2\gamma_C^2}\right), \qquad (5)$$

where c_p is the audio-visual coherence $c(\mathbf{x})$ corresponding to pixel p with spatio-temporal coordinates $\mathbf{x}$. We fix $\gamma_C = 0.1$ to assign a low weight to links between neighboring pixels with different coherence. Since in this case $C_{p,q} \neq C_{q,p}$ if $c_p \neq c_q$, our graph is directed. The proposed audio-visual term is thus similar to the boundary term in the sense that it is computed between neighboring pixels. Furthermore, low weights are assigned to the edges that link pixels belonging to different regions (in this case regions presenting high and low coherence instead of regions with significantly different color). However, our audio-visual term does not affect regions with low audio-visual coherence. Notice that the weight $C_{p,q}$ is directly proportional to the audio-visual coherence in the origin pixel c_p and thus the weight of the links is close to zero in regions with low coherence. Thus, our audio-visual term links together only neighboring points that present a similar and *relevant* audio-visual coherence. This represents the main difference between our *audio-visual term* and the term in [5, 6]. In their case, all the pixels are linked to the background and foreground terminals according to their audio-visual correlation. Thus, when a part of the audio-visual object has a low coherence with the audio signal, the segmentation process assigns this part to the background. For example, some applications such as the speaker's face extraction might be interested in extracting the speaker's forehand even though it does not present a high coherence with the speech. Thus, our term links together neighboring regions with high audio-visual coherence without penalizing or making any assumptions about the remaining video regions.

4. AUDIO-VISUAL SEGMENTATION PRIORS

The segmentation procedure presented in the previous section requires an starting point for the segmentation process, i.e. some initial information about the foreground (audio-visual object) and background location. As explained before, this prior information is obtained from the fusion of audio and video modalities. From Sec. 2 we can extract the pixels that are likely to compose the audio-visual object, that are those pixels that have a high audio-visual coherence.

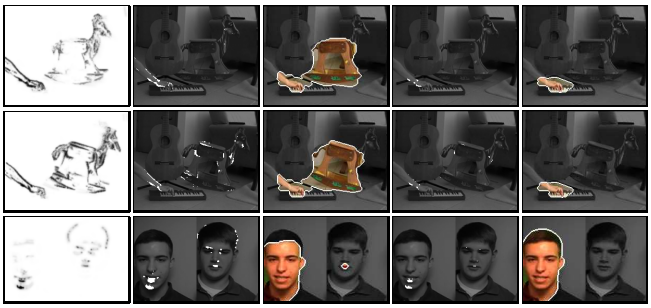
Let P be the number of pixels in the video GoF. The number of seeds that are automatically chosen for foreground N_f and background N_b are $N_m = PH_m$ for $m = \{f, b\}$, where the quantities H_f and H_b can be fixed depending on the application. The foreground seeds are chosen to be the N_f pixels with highest audio-visual coherence c_p, while the N_b constraints for the background are *randomly* distributed in the GoF. This election ensures that no additional assumptions are made. In [5, 6] the pixels presenting a low audio-visual correlation were assumed to belong to the background and thus they could not be included in the extracted region. In all experiments we use $H_f = H_b = 3 \cdot 10^{-3}$, i.e. a 0.3% of the pixels are automatically assigned to foreground and background. This value is low because we want to be sure to introduce the smallest possible number of errors in the initial labeling. A choice of $H_f > H_b$ can lead to the extraction of a larger region.

In our work, no segmentation seeds are fixed in the video frames in silent periods. Since in this frames the audio-visual coherence is very low, no seeds would be fixed for the foreground and the introduction of background constraints would only penalize the extraction of the audio-visual object. Since the seeds choice is unsupervised we fix the weight that links the seeds to the corresponding terminal (F or B) to the maximum weight of a n-link: $W = \max_{p \in \mathcal{P}}(V_{p,q} + \lambda_C C_{p,q})$. This value is high enough to influence the segmentation but the label can be modified by the min-cut max-flow algorithm if required (for example when a foreground constraint is isolated in the background).

5. EXPERIMENTS

We test the proposed audio-visual segmentation algorithm in fragments of sequences containing non-stationary sound sources and distracting moving objects. Each video fragment is around 1 second length (the GoFs are composed by $N_t = 25$ frames). In all experiments the parameters are fixed as suggested in Sec. 3.

Fig. 3 shows the results when analyzing two sequences containing a strong distracting motion. The first clip is taken from the state-of- the-art source localization work presented by Kidron et al. in [3], and it features a hand playing a synthesizer (non-stationary sound source) and a wooden rocking horse is moving in the background. The second sequence is a synthetic sequence composed of a



a) Coherence $c(\mathbf{x})$ b) Source seeds - Result $H_b = H_f = 10 \cdot 10^{-3}$ c) Source seeds - Result $H_b = H_f = 3 \cdot 10^{-3}$

Fig. 3. Extracted audio-visual objects in sequences containing distracting video motion for a different number of initial seeds.

fragment of clips g01 and g08 from the groups partition of CUAVE database [11] in which two persons are present: the left person is uttering some numbers and the right one is mouthing the same numbers. Thus, both sequences are composed of a moving object associated to the audio signal (hand and left person) and another one that represents a strong visual distraction, whose motion is either periodic [top] or very similar to the motion in the audio-visual object [bottom]. When using a very small number of seeds (c) as suggested in Sec. 4, the audio-visual object is successfully determined for both clips and the extracted region does not contain the distracting moving objects. In this case, few labels are wrong (located over the horse or the wrong person) because only the 0.3% of pixels in the GoF are initially labelled. However, when H_f and H_b increase drastically (in (b) we have 1% of seeds), the number of foreground seeds located in the distracting moving objects grows too and the extracted region can contain parts that do not belong to the audio-visual object.

Figure 4 shows a comparison between the extracted audio-visual objects obtained with our method [bottom] and the methods in [5, 6] [top] when analyzing sequences g22 and g23 of CUAVE database. Our results are specially favorable in (c): the region that we extract contains the complete mouth region while in [5] it was mostly composed of the girl's hair. In (e) our approach extracts completely the girl's face because the presence of the regional term makes easier the extraction of regions homogeneous in color. In contrast, only the mouth region can be extracted in [6] [top] because their audio-visual term penalizes pixels presenting a low coherence with the soundtrack. In Figure 4 we can also compare the results *with* [bottom row] and *without* [third row] the audio-visual term in equation (2). In (a) and (c), when $\lambda_C = 0$ the current speaker's mouth region is only partially extracted. The introduction of the proposed audio-visual term links together the pixels in the speaker's mouth since in this region the audio-visual coherence is high. As a result, the label of the seeds is efficiently spread and the complete mouth region is extracted.

6. DISCUSSION

We have presented a novel method which is able to automatically extract the audio-visual objects present in a scene. Our approach has been tested in challenging sequences containing distracting motion and non-stationary sound sources. In all cases the video modality of the sound source has been successfully extracted. Our definition of the segmentation problem, which includes an audio-visual term *and* a regional term encouraging homogeneous regions, makes our method suitable for applications that require the extraction of the complete audio-visual object. For example the whole speaker's face region might be needed when trying to protect the speaker's identity by automatically mosaicing his face.

(a) (b) (c) (d) (e)

Fig. 4. [From top to bottom] Extracted regions when applying the method in [5] to sequence g23 [left] and the approach in [6] to movie g22 [right]; Foreground seeds chosen using the audio-visual coherence; Results when the audio-visual term is not used ($\lambda_C = 0$); Our results when both audio-visual and regional terms are considered. In all situations the current speaker is detected.

7. REFERENCES

[1] J. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *NIPS*, 1999.

[2] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[3] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. on Signal Processing*, vol. 55, no. 4, pp. 1390–1404, 2007.

[4] A. Llagostera Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind Audio-Visual Source Separation based on Sparse Redundant Representations," *IEEE Trans. on Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.

[5] Y. Liu and Y. Sato, "Finding Speaker Face Region by Audio-visual Correlation," in *Workshop on M2SFA2*, 2008.

[6] Y. Liu and Y. Sato, "Visual localization of non-stationary sound sources," in *MM '09: Proc. of ACM int. conf. on Multimedia*, 2009.

[7] Y. Boykov and M. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. of IEEE ICCV*, 2001.

[8] Y. Li, J. Sun, and H. Shum, "Video object cut and paste," *Proc. of ACM SIGGRAPH*, 2005.

[9] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *Proc. of ACM SIGGRAPH*, 2004.

[10] A. Llagostera Casanovas and P. Vandergheynst, "Nonlinear Video Diffusion based on Audio-Video Synchrony," *IEEE Trans. on Multimedia* (submitted to), [Online] Available: http://infoscience.epfl.ch/record/151692/files/, 2010.

[11] 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, Nov. 2002.