Separating Transparent Layers of Repetitive Dynamic Behaviors

Bernard Sarel Michal Irani

Dept. of Computer Science and Applied Math.
The Weizmann Institute of Science
76100 Rehovot, Israel

Abstract

In this paper we present an approach for separating two transparent layers of complex non-rigid scene dynamics. The dynamics in one of the layers is assumed to be repetitive, while the other can have any arbitrary dynamics. Such repetitive dynamics includes, among other, human actions in video (e.g., a walking person), or a repetitive musical tune in audio signals. We use a global-to-local space-time alignment approach to detect and align the repetitive behavior. Once aligned, a median operator applied to space-time derivatives is used to recover the intrinsic repeating behavior, and separate it from the other transparent layer. We show results on synthetic and real video sequences. In addition, we show the applicability of our approach to separating mixed audio signals (from a single source).

1. Introduction

Our urban environment is full of transparent surfaces which induce images and videos with mixed layers. The separation of transparent layers serves as a pre-processing stage for many vision algorithms which face grave difficulties in the presence of a superimposed layer, such as recognition, segmentation, feature extraction, etc. Separating transparent layers from a single video sequence when the two layers contain complex non-rigid motions is a very difficult problem.

Previous approaches for layer separation in video assume that dense correspondences can be pre-computed for each pixel in each layer across the entire sequence [2, 4, 8, 9, 12]. These methods are therefore mostly restricted to scenes with simple 2D parametric motions, which are easy to compute under transparency and provide dense correspondences.

Estimating frame-to-frame correspondences between successive image frames in the case of complex non-rigid motions is not a reliable process. This problem becomes even worse when the non-rigid motion is superimposed by yet another transparent layer with a different complex non-rigid motion. Thus, existing approaches to layer separation cannot be applied to such video sequences, since they rely on accurate recovery of the frame-to-frame correspondences between successive frames (in both layers). The case of one transparent layer having arbitrary non-rigid motion was addressed by [7], yet to achieve this, the other layer was assumed to be characterized by a 2D parametric transformation. None of the existing methods can handle *two dynamic* layers.

In this paper we propose an approach for separating transparent layers where one of the layers includes *repetitive* dynamics (which may be non-rigid and complex). Repetitive behavior is very common in the natural world, e.g., human and animal actions, repetitive tunes in audio, etc. The other layer is unrestricted and may include arbitrary non-rigid dynamics. We demonstrate the applicability of our approach, both in video and in audio.

The repetitiveness in one of the layers facilitates its detection. All segments that include the repetitive behavior, are very similar. By definition, these segments differ due to the other superimposed transparent layer. But moreover, they differ in the repetitive behavior as well, both globally (on the space-time scale of the whole repetition), and locally (on the scale of small space-time behavioral details). Consider for example a person walking. The person may be walking with different speeds, or have different gait in each repetition. Therefore a global-to-local approach is used. *Global space-time alignment* of the repetitive behavior compensates for the differences resulting from different camera distances, different zooms, and different speeds of the intrinsic behavior in its different occurrences. Local space-time fluctuations of the repetitive behavior, necessitate an additional *local space-time refinement* stage afterwards. This produces for each space-time position a set of corresponding space-time points across all repetitions. Note that such correspondences cannot be obtained using optical flow estimators, due to the highly non-rigid complex motions in the sequence. Once the different occurrences of the intrinsic behavior are brought into alignment, a median is applied to the space-time derivatives to separate the two non-rigid layers.

The rest of the paper is organized as follows. In Section 2 we give an overview of our approach. In Section 3 we describe the use of global space-time alignment for finding repetitions and the global-to-local approach for finding sets of corresponding points. In Sections 4 and 5 we describe the recovery of the first and second layers, respectively, and show results on video sequences. In Section 6 we show the applicability of our approach to other domains by applying it to audio signals.

2. Overview of Our Approach

We consider the general case of two superimposed transparent space-time layers in a video or audio sequence, where one layer has complex non-rigid deformations over time pertaining to repetitive dynamic behavior (e.g., a walking in video, or a tune that repeats itself in audio). The second layer can have any non-rigid deformations over time. We take advantage of the repetitive nature of the dynamic behavior of one of the layers in order to achieve layer separation.

Natural non-rigid repetitive actions, such as a person walking, or running, which cannot be modelled by simple 2D parametric transformations between consecutive video frames, induce complex non-rigid motion fields. No optical flow estimator would be able to recover the frame-to-frame correspondences between successive frames in such complex sequences. Yet if we treat the entire space-time volume of a video segment containing a single repetition of the dynamic behavior, as a single unit, then a simple volumetric *space-time* parametric transformation captures the global spatio-temporal deformations between multiple occurrences of the same dynamic behavior in the video sequence.

This notion of repetition of space-time behavior characteristics is key to our approach. First, we automatically detect the temporal extent and the temporal distances between the different occurrences of the intrinsic repetitive dynamic phenomenon in the video sequence (or in the audio signal). Then we use a global space-time alignment procedure applied to the video segments where the repetitive phenomenon was detected. This accounts for global changes in space and time between the various repetitions (such as different distance from the camera, differences in speed of the performed action, etc.), and therefore brings the input video segments containing the "intrinsic" repetitive behavior into *coarse alignment*. An additional spatio-temporal *local refinement* is then computed, to account for small local deformations in the way the repetitive behavior is manifested locally in space-time (such as raising an arm to a different height in different repetitions of the action).

Once perfect space-time alignment of the repetitive behavior has been obtained, we can proceed to extract and separate it from the other superimposed layer. This is done by extending the intrinsic image-based approach of Weiss [11] to space-time volume. A median is applied to the spatio-temporal derivatives of all corresponding space-time points of the "intrinsic dynamic behavior" recovered by the global-to-local alignment procedure. This operation removes the derivatives of the other arbitrary layer.

Lastly, we apply the "Layer Information Exchange" algorithm of [7] between the original movie sequence and the first recovered layer, thus removing the first layer from the original movie and obtaining the second separated transparent layer.

3. Detecting Repetitive Behaviors

To detect and align multiple occurrences of the same dynamic behavior in the video sequence we use global-to-local space-time alignment.

3.1. Global Space-Time Alignment

For the global space-time alignment we use the approach of Ukrainitz and Irani [10], which extends the image alignment method of Irani and Anandan [3] into space-time. This approach is useful for locking onto the dominant space-time parametric transformation (in our case – the repetitive behavior), despite the presence of the other superimposed transparent layer.

Formally, given a video sequence S, and two arbitrary space-time segments A and B in S, we seek the spatio-temporal parametric transformation $\vec{p}$ that maximizes a global similarity measure M between these two segments after bringing them into alignment according to $\vec{p}$. Fig. 1 graphically illustrates this step. The volumetric space-time parametric transformation, $\vec{p}$, comprises of a *2D affine transformation in space* (accounting for differences in zoom and orientation between the different occurrences of the intrinsic dynamic behavior), and a *1D affine transformation in time* (accounting for differences in speed of the intrinsic behavior in the different occurrences). Therefore, the space-time transformation $\vec{p}$ comprises of 8 parameters, where the first 6 parameters $(p_1, \ldots, p_6)$ capture the spatial 2D affine transformation and the remaining 2 parameters (p_7, p_8) capture the temporal 1D affine transformation. The spatio-temporal displacement vector $\vec{u}(x, y, t; \vec{p})$ is therefore:

$$\vec{u}(x,y,t;\vec{p}) = \begin{bmatrix} u_1(x,y,t;\vec{p}) \\ u_2(x,y,t;\vec{p}) \\ u_3(x,y,t;\vec{p}) \end{bmatrix} = \begin{bmatrix} p_1 x + p_2 y + p_3 \\ p_4 x + p_5 y + p_6 \\ p_7 t + p_8 \end{bmatrix} \quad (1)$$

To make the paper self-contained, we will briefly review the approach of [10].

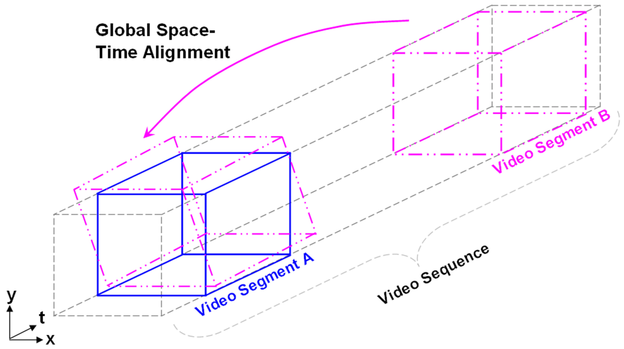

Figure 1: **Global Space-Time Alignment**
Movie segment B in the video sequence is space-time aligned with movie segment A using a space-time volumetric parametric transformation.

3.1.1 The Similarity Measure

Local normalized correlations are computed within small *space-time patches* (e.g. $7 \times 7 \times 7$). The *global* similarity measure M is then computed as the sum of all those local measures in the entire sequence segment. The resulting global similarity measure is thus invariant to spatially and temporally varying non-linear intensity transformations, which in our case may result from the superposition of transparent layers.

Given two corresponding space-time patches/windows, w_A and w_B, one from each sequence segment, their local Normalized Correlation (NC) can be estimated as $NC(w_A, w_B) = \frac{\text{cov}(w_A,w_B)}{\sqrt{\text{var}(w_A)}\sqrt{\text{var}(w_B)}}$, where cov and var stand for the covariance and variance of intensities. Squaring the NC measure further accounts for correlation of information in cases of contrast reversal as well. The patchwise local similarity measure is therefore:

$$C(w_A, w_B) = \frac{\text{cov}^2(w_A, w_B)}{\text{var}(w_A)\text{var}(w_B) + \alpha} \qquad (2)$$

where the constant α is added to account for noise (we used $\alpha = 10$, but the algorithm is not particularly sensitive to the choice of α).

The *global* similarity measure M between two sequence segments A and B is computed as the sum of all the *local* measures C applied to small space-time patches around each pixel in the sequence:

$$M(A, B) = \sum_x \sum_y \sum_t C\left(w_A(x, y, t), w_B(x, y, t)\right) \qquad (3)$$

This results in a global measure which is invariant to highly non-linear intensity transformations (which may vary spatially and temporally within and across the sequence segments).

3.1.2 Alignment Algorithm

Our goal is to recover the global geometric space-time transformation which maximizes the global measure M between the two sequence segments A and B. The local measure C and the global measure M can be expressed in terms of the unknown parametric transformation $\vec{p}$:

$$M(\vec{p}) = \sum_{(x,y,t)\in A} C\left(w_A(x, y, t), w_B(x + u_1, y + u_2, t + u_3)\right) \qquad (4)$$

(recall that the spatio-temporal displacement vector $\vec{u}$ depends on $\vec{p}$ – see Eq. (1)). The alignment problem becomes finding the spatio-temporal transformation $\vec{p}$ which maximizes the global similarity measure $M(\vec{p})$.

For the optimization task the Newton method is used. $M(\vec{p})$ is locally approximated quadratically around the current spatio-temporal transformation estimate $\vec{p}_0$. The spatio-temporal update step $\vec{\delta}_p = \vec{p} - \vec{p}_0$ is found by differentiating the local quadratic approximation with respect to it, and equating to zero. The resulting update step is:

$$\vec{\delta}_p = -\left(H_M(\vec{p}_0)\right)^{-1} \cdot \nabla_{\vec{p}} M(\vec{p}_0) \qquad (5)$$

The Hessian H_M and gradient $\nabla_{\vec{p}} M$ are evaluated at each space-time point using the gradient and Hessian of C (through Eq. (4)). For more details see [10].

This procedure is performed using a space-time Gaussian pyramid for both sequence segments A and B. Several maximization iterations are performed in each level until convergence, and the result is used for initializing the next pyramid level.

3.2. Detecting Repetition Length and Extent

The nature of the repetitive non-rigid dynamic behavior is not known in advance, and we would like to automatically detect the repetition length, the number of repetitions, and their positions in the movie. We choose an *arbitrary* segment of the movie (a space-time volume – typically a few tens of entire frames) to be our reference "test meter" for repetitions. We compare it to a "sliding space-time window" of the same size, starting at the beginning of the movie. This is done by bringing into best possible space-time alignment the reference segment and the current transformed sliding window using the procedure in Section 3.1. We then compute their "degree of similarity" using the *global* similarity measure M in Eq. (3). This process associates with each point in time the degree of similarity between its surrounding space-time video segment to the reference segment (after accounting for global deformations in space and in time between these two video segments).

Fig. 2.a displays the results of such measurements of a real video sequence as a function of time (frame-number),

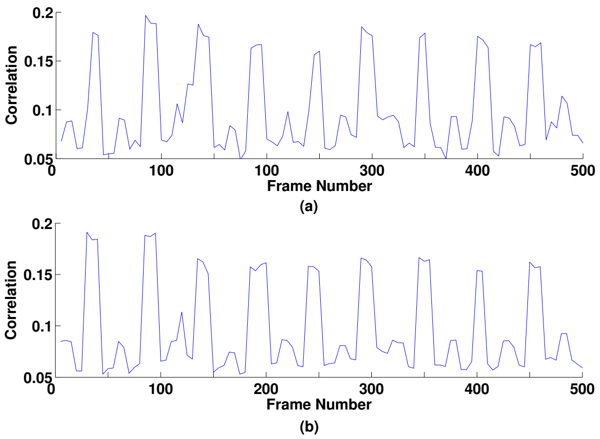

Figure 2: **Finding Dynamic Behavior Repetitions**
The graphs display the correlation measure of a reference segment of a real video sequence, to a "sliding window" with the same number of frames in the movie. The position of the "sliding window" in the sequence is marked by its first frame (X-axis). (a) The case where the reference segment is 20 frames long (shorter than the average repetition length). (b) The case where the reference segment is 60 frames long (longer than the average repetition length).

where the reference video segment was much shorter than the average repetition length. Fig. 2.b displays the resulting measurements for the same sequence, but this time with a reference segment larger than the average repetition length. As evident from Fig. 2.a and Fig. 2.b, the repetition of the non-rigid dynamic behavior is exhibited clearly. Moreover, qualitatively and quantitatively there is no substantial difference in using different reference segment lengths, making this a robust method for detection of repetitive phenomena.

The actual repetition length and the number of repetitions can be easily extracted from the measurement graph. Each repetition starting and ending point in the video sequence can be found automatically (e.g., by identifying sharp rises in the graph).

Note that trying to detect these repetitions with a single image frame (e.g., by trying to correlate it to all other frames in the sequence after best image-to-image alignment) fails for two reasons: (i) Different occurrences of the intrinsic action have temporal sub-frame misalignments between them (e.g., due to differences in speed), which means that their individual frames do not capture exactly the same body poses (namely, they sample the action at different time instances). Such temporal sub-frame misalignments are recovered and compensated for by the global sequence-to-sequences alignment process, but are not handled by image-to-image alignment. (ii) An entire space-time volume contains significantly more information than a single image, thus allowing to lock onto the dominant intrinsic action, whereas the spatial information alone in a single snapshot in time (in a single frame) is not salient enough.

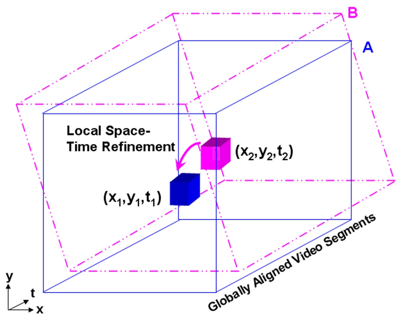

Figure 3: **Local Space-Time Alignment Refinement**

After global parametric alignment of the video segments A and B, there are residual (non-parametric) misalignments. For each point (x_1, y_1, t_1) in A, the local refinement step seeks a better local match (x_2, y_2, t_2) in B, within a small space-time neighborhood around the corresponding point induced by the global alignment.

3.3. Local Space-Time Refinement

Real non-rigid behavior is never perfectly repetitive. A person for example, does not repeat the phases of walking or running exactly in the same fashion . On top of the global variations (e.g., changing speed, position, or orientation) which are already accounted for by the global sequence-to-sequence alignment, there are also local deformations in the dynamic behavior (e.g., small changes in the relative positioning of body parts in the same motion phase). This results in *local residual misalignments* both in space and in time. Therefore, a local refinement procedure for the alignment of the space-time features in necessary.

For each space-time point we seek a better local match within the other segments aligned to it. This is done by correlating a small space-time patch around the point (typically 7x7x7), with nearby patches in the *other* globally aligned segments (typically up to spatio-temporal displacements of ± 4 pixels spatially and ± 2 frames temporarily). This is done using the correlation measure C of Eq. (2).

Fig. 3 graphically displays the local space-time alignment step of a small space-time patch centered at (x_1, y_1, t_1) in one video segment, to its best correlation match in a small neighborhood in another video segment after global space-time alignment.

Fig. 4.a shows a few frames of a real video sequence of two transparent layers. The layer containing the repetitive dynamics comprises of a jumping man filmed through a transparent swivelling door, while the other arbitrary layer is the background reflected in the door's glass window (peo-

ple playing basketball). Fig. 4.b, shows the corresponding frames for each image in Fig. 4.a in some other movie segments, detected via global alignment. In Fig. 4.c we overlayed the images from Fig. 4.a and Fig. 4.b so that the residual misalignment of the man's body after the global alignment becomes visually evident. In Fig. 4.d we overlayed the images from Fig. 4.a with the corresponding images from Fig. 4.b after local refinement (seeking best local space-time correspondence for each point in the images in Fig. 4.a, in a small space-time *neighborhood* of the points in the images in Fig. 4.b), so that the full alignment of the man's body after the local refinement becomes visually evident.

4. Extracting the First Layer

The separation of the two space-time transparent layers is performed in the domain of the spatio-temporal derivatives (S_x, S_y, S_t), where S is the input video sequence. To isolate the non-rigid intrinsic behavior and separate it from the other transparent layer we apply the median operator to the spatio-temporal derivatives of each space-time point (x, y, t) and its multiple correspondences, after having brought into global and local alignment multiple occurrences of the intrinsic dynamic behavior.

The median, in effect, removes the derivatives of the transient dynamics, i.e., the non-rigid motion of the other layer, since these are *not* aligned (see [11, 7]). The only remaining non-zero derivatives are of the repetitive dynamic intrinsic behavior. Integrating the resulting derivatives yields the gray level sequence of the intrinsic behavior repeated for all its occurrences. At each time a different behavior occurrence serves as a reference coordinate system, with its global and local variations, thus creating the full sequence of the first transparent layer. Such an example can be found in Fig. 5.b.

5. Extracting the Second Layer

To extract the other non-rigid *arbitrary* transparent layer, we use the algorithm proposed by [7] for "Layer Information Exchange". The first recovered layer L_1, is subtracted from the input sequence S to obtain a second layer $L_2 = S - \alpha L_1$ (where α is a scalar). We seek α which minimizes the correlation between L_1 and L_2. Such an α provides the best layer separation. Since α is unlikely to be uniform in the entire sequence, we currently compute a local α for each pixel according to its local surrounding neighborhood. This local separation procedure is applied to the derivatives of the sequence (separately on each directional derivative). The second layer is then recovered by integration.

Results of such layer separation can be found in Fig. 5. Fig. 5.a displays three frames from the same video sequence

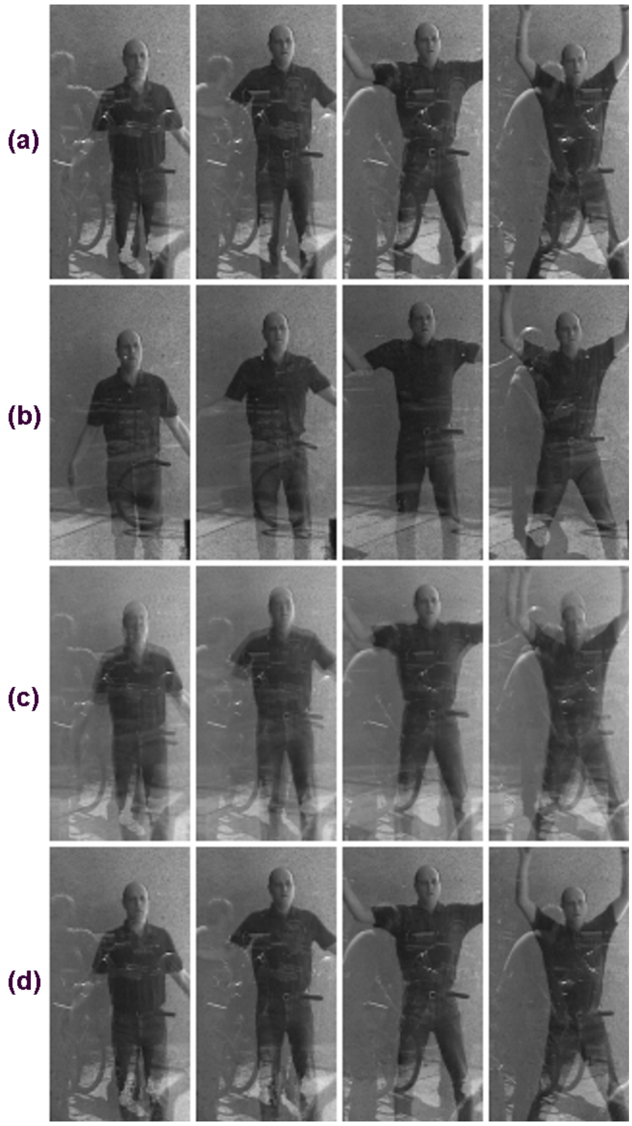

Figure 4: **Global-to-local Space-Time Alignment**
The global-to-local alignment process is displayed. (a) Four example frames from the movie. (b) Their corresponding frames in some other repetitions. (c) Residual misalignment of (a) and (b) after applying **global** *space-time alignment. (d) No residual misalignment after applying* **local** *space-time refinement. The video sequences can be viewed at* **www.wisdom.weizmann.ac.il/$\sim$vision/RepetitiveTransp.html**

as in Fig. 4 (a jumping man filmed through a glass door and other people reflected in the same door). Fig. 5.b displays the first extracted layer, and Fig. 5.c displays the second extracted layer.

Another example is displayed in Fig. 6. In this case we manually mixed two video sequences. We took a video of a walking man on a uniform background and superimposed it on another video of running water in a small garden creek

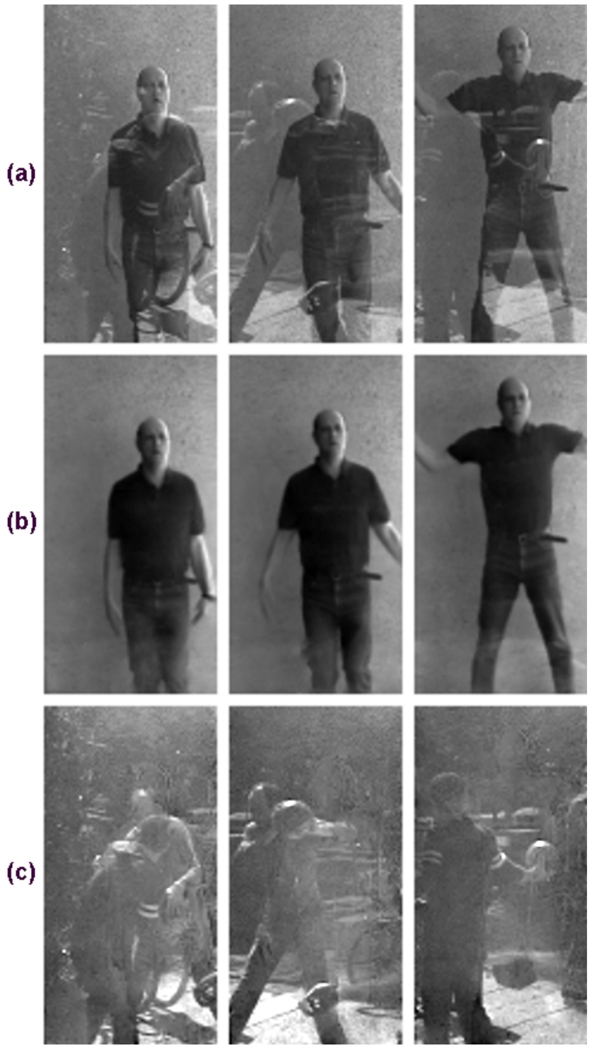

Figure 5: **Layer Separation in Real Video Transparency**
(a) Three frames from a real video of a jumping man seen through a swivelling glass door, while other people playing basketball are reflected in the glass door. (b) The first recovered layer – the jumping man. (c) The second recovered layer – the basketball players. The video sequences can be viewed at **www.wisdom.weizmann.ac.il/∼vision/RepetitiveTransp.html**

Figure 6: **Layer Separation in Video**
(a) Three frames from the input video (which was composed of two real videos superimposed: a man walking, and a waterfall in a small creek). (b) The first recovered layer – the walking man. (c) The second recovered layer – the waterfall in the creek. The video sequences can be viewed at **www.wisdom.weizmann.ac.il/∼vision/RepetitiveTransp.html**

(a highly non-rigid dynamic scene). Fig. 6.a displays three frames from the superimposed video. Fig. 6.b displays the first recovered layer (the walking man), and Fig. 6.c displays the second recovered layer (the running water).

6. Separating Mixed Audio

Our approach is not restricted to video sequences, and can be applied to repetitive behavior found in other domains. We show its applicability for separating a repetitive tune superimposed with some other audio signal. A segment of the masterpiece "Bolero" by Ravel was repeated a number of times and mixed with a recording of a man talking. In another instance it was mixed with another song. Similar methods as applied to video sequences are applied here as well.

Fig. 7.a displays the time-frequency representation of "Bolero" mixed with the talking man. It displays the discreet time Fourier transform of the audio track, at small consecutive time intervals. The horizontal axis represents time, and the vertical axis represents the frequencies present in the signal in a small time window centered at that particular time instance. As the signal changes over time, different frequencies are dominant. It is evident from Fig. 7.a that the frequency content has a repeating pattern, which can be detected and used for finding the repeating segments of the audio track. Global alignment can be achieved using image-to-image alignment methods (e.g., [2, 4]) applied to the time-frequency data (treating it as an image) with a sliding window in time. Fig. 7.b displays the correlation of a reference segment from the time-frequency domain in Fig. 7.a, with a sliding window of the same size within the domain. The peaks in Fig. 7.b match the position of repetitions and their extent within Fig. 7.a (we also tried finding those repetitions in the 1D raw audio signal, but this failed to provide any meaningful information about the repetitions as opposed to the above 2D image-based approach). The layer separation of Sections 4 and 5 can then be used for

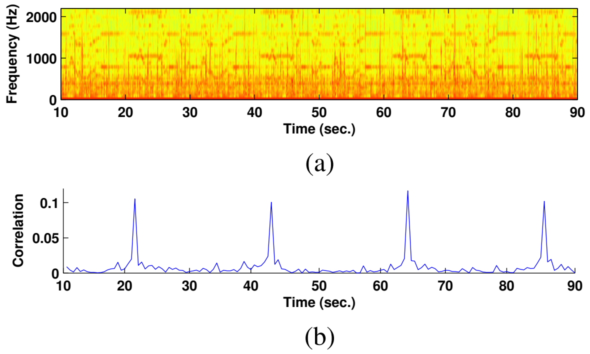

Figure 7: **Finding Repetitions in Mixed Audio**
(a) This is a time-frequency domain spectrogram of a superposition between the recording of a talking man and a repetitive tune from "Bolero" by Ravel. The four repetitive segments of the intrinsic phenomena are evident in the horizontal bars representing energy in some specific frequency bands over time. (b) Detected repetitions using image alignment approach by aligning a short segment (5 secs.) against a "sliding window" in (a). The repetitions in the first layer are detected via the sharp peaks. The audio files can be accessed at **www.wisdom.weizmann.ac.il/∼vision/RepetitiveTransp.html**

separating the two signals.

Standard Independent Component Analysis (ICA) methods [1] cannot be used in this case since they require the same number of inputs as the number of layers. Previous methods for single source audio separation [5, 6] have assumed that the time-frequency domain is either strictly divided between the two sources [6] (i.e., the frequency bands in small time-frequency windows belong to just one source or the other, but not to both), or allowed weighted superposition [5]. They rely on a training phase for learning the signal time-frequency characteristics in order to separate the signals. Here the separation is obtained from a single input source without previous knowledge of the nature of the source, and with no learning phase.

7. Summary

In this paper we present an approach for separating two transparent layers of complex non-rigid scene dynamics. The dynamics in one of the layers is assumed to be repetitive, while the other can have any arbitrary dynamics. Such repetitive dynamics includes, among other, human actions in video (e.g., a walking person), or a repetitive musical tune in audio signals. We show the applicability of our approach to separating video sequences and audio signals (from a single source).

8. Acknowledgements

This research was conducted in the Moross Laboratory for Vision and Motor Control at the Weizmann Institute of Science.

References

[1] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[2] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding, CVIU*, 63(1):75–104, January 1996.

[3] M. Irani and P. Anandan. Robust multi-sensor image alignment. pages 959–966. ICCV, January 1998.

[4] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, December 1993.

[5] G.-J. Jang, T.-W. Lee, and Y.-H. Oh. Single channel separation using map-based subspace decomposition. *Electronic Letters*, 39(24), November 2003.

[6] S. T. Roweis. One microphone source separation. pages 793–799. NIPS 13, December 2001.

[7] B. Sarel and M. Irani. Separating transparent layers through layer information exchange. volume 4, pages 328–341. ECCV, May 2004.

[8] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. pages 246–253. CVPR, June 2000.

[9] Y. Tsin, S. B. Kang, and R. Szeliski. Stereo matching with reflections and translucency. volume I, pages 702–709. CVPR, June 2003.

[10] Y. Ukrainitz and M. Irani. Aligning sequences and actions by maximizing space-time correlations. Technical Report MCS05-06, (can be accessed at: wisdomarchive.wisdom.weizmann.ac.il:81/archive/00000377/), Weizmann Institute of Science, 2005.

[11] Y. Weiss. Deriving intrinsic images from image sequences. pages 68–75. ICCV, July 2001.

[12] Y. Wexler, A. Fitzgibbon, and A. Zisserman. Bayesian estimation of layers from multiple images. pages 487–501. ECCV, May 2002.