Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns

Yong Rui and P. Anandan
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
E-mail: {yongrui, anandan}@microsoft.com

Abstract

The analysis of human action captured in video sequences has been a topic of considerable interest in computer vision. Much of the previous work has focused on the problem of action or activity recognition, but ignored the problem of detecting action boundaries in a video sequence containing unfamiliar and arbitrary visual actions. This paper presents an approach to this problem based on detecting temporal *discontinuities of the* spatial pattern *of image motion that captures the action. We represent frame to frame optical-flow in terms of the coefficients of the most significant principal components computed from all the flow-fields within a given video sequence. We then detect the discontinuities in the temporal trajectories of these coefficients based on three different measures. We compare our segment boundaries against those detected by human observers on the same sequences in a recent independent psychological study of human perception of visual events. We show experimental results on the two sequences that were used in this study. Our experimental results are promising both from visual evaluation and when compared against the results of the psychological study.*

1. Introduction

1.1. Background and motivation

The analysis of human action captured in video sequences has been a topic of considerable interest in computer vision. The primary motivation has been the recognition of human action and activities (e.g., see [1, 2, 3] just to cite a few examples). Most researchers have preferred a strongly model-based approach – i.e., they have either focused on a few specific actions (e.g., [1, 2]) or relied heavily on the scene and situational context to detect particular pattern of activities (e.g, see [3] for such an effort in the context of aerobics exercise). For the problem of recognition and detection of familiar actions and activities, this is perhaps the most sensible approach one can take.

In this paper, we are interested in a slightly different problem, namely that of segmenting a long *continuous* video sequence into short segments that correspond to smooth or continuous actions or sub-components of actions. While we believe that as in the case of static image segmentation, video action segmentation is likely to be most reliable and accurate when it is closely coupled with recognition (i.e., identify familiar actions), the number of actions for which we actually have visual models is quite limited. On the other hand, since visual actions are captured directly by image motion, temporal discontinuity of image motion may be expected to be a strong cue for possible places to segment the video.

These two types of segmentation roughly correspond to a high level top-down analysis of the video data and a bottom-up image based analysis. For example, considering a video sequence in which a subject is performing a common activity, such as making a bed, a high level segmentation would be in terms of activities such as "pulling the sheet", "setting the pillows", etc. However, each of these actions themselves involves movement of the subject's body, arms and legs in one direction or the other. The low level segmentation would be at the end of a continuous visual motion (such as moving to the left, swinging the arms in one direction, etc.). Our work focuses on this level, namely that of segmenting a continuous video sequence into smaller segments based on discontinuity in the *pattern* of image motion. Since the video frames corresponding to these segment boundaries capture important human action (posture) changes, we will refer these frames as *keypose* frames. In this paper, we use *keyposes* and action segment boundaries interchangeably.

We are motivated to work on this problem for multiple reasons. First, we are interested in creating a compact summary of the activity, which when viewed in a continuous fashion conveys the impression of the action in the input sequence. Second, the low-level segment boundaries will, in general, be a superset of the high-level segment bound-

aries; hence, these are building blocks for semantics-based segmentation. Third, since we are detecting discontinuity in smooth image motion, the segment boundaries we detect will be good candidates for "I-frames" in a compression strategy [4].

It is important to distinguish our work from the more traditional work on scene cut detection [5]. Typically the scene cut detection techniques detect shot-boundaries – i.e., the places that mark the end of a continuous video shot (either due to shutting off the camera or due to editing in a different camera sequence), whereas we are interested in segmentation within a single "shot".

It is also interesting to note that in psychological studies [6, 7] human subjects, when asked to perform "fine-grained" segmentation, appear to choose locations in time that are based on image domain motion information. On the other hand, coarse-grain segmentation by the same subjects appears to be more influenced by top-down and contextual analysis. We obtained the data used in the psychological studies by Zacks, *et al.* [6] and compared our segmentation results with the fine-grained segmentation performed by human subjects. This paper includes the results of this comparison.

1.2. A Summary of our approach

Our basic premise is that the *temporal* segment boundaries correspond to *temporal* discontinuity in the *spatial pattern* of image motion. The key here is that we want to measure the discontinuity of the entire spatial pattern of motion of the subject in the video, not the motion vector of any particular pixel. In order to capture the pattern of motion, we performed a Singular Value Decomposition (SVD) of the set of frame-to-frame optical flow fields collected over an entire video clip. After removing the low energy basis vectors (in order to reduce noise), we used the coefficients of the remaining basis vectors as the representation of the spatial motion pattern. We analyzed the temporal trajectories of these coefficients, and detected discontinuities in them. In this paper, we compare three different measures of discontinuity on the SVD coefficients.

The rest of the paper is organized as follows. A detailed description of our algorithm is given in Section 2. Section 3 describes the results of applying our algorithm to two video sequences, also used in the psychological studies by Zacks. Section 4 summarizes the paper and discusses the possible directions for future research.

2. The Proposed Algorithm

The proposed algorithm consists of several stages.

1. We first compute the frame to frame dense optical flow for all the frames in the sequence. This is described in Section 2.1.
2. Foreground object segmentation maps are then used as the masks to get rid of irrelevant (and noisy) flow vectors in the static background. This stage is described in Section 2.2.
3. Each flow field is represented as a long vector and the SVD of the collection of all the flow fields is computed. This determines a set of "basis" flow fields. For each flow field, only the coefficients associated with the most significant basis flows are used. The remaining coefficients are ignored. This provides a degree of noise reduction as well as a more compact set of parameters to use for further analysis. This is described in detail in Section 2.3.
4. The temporal trajectories of the SVD coefficients of the flow field are analyzed to determine locations of the segment boundaries. We have explored three different methods for temporal discontinuity detection. These are described in Section 2.4.

A flow chart of the complete algorithm is shown in Figure 1.

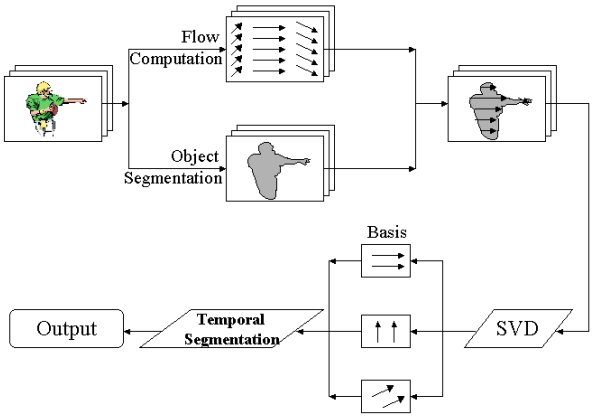


Figure 1. Flow chart of the algorithm.

2.1. Frame to frame optical flow computation

Since our entire analysis is based on the spatio-temporal pattern of image motion, the first step is to compute the image flow fields. We use the hierarchical optical flow algorithm of Szeliski [8] which combines many of the best features of existing optical flow techniques. In particular, it incorporates multi-resolution estimation, regularization in the form of a spatial smoothness constraint, and combines correlation matching together with "Lucas-Kanade" type gradient-based refinement. Although Szeliski's algorithm allows for the use of temporal smoothness constraints, since we are interested in detecting temporal motion discontinuities, we did not use this option. We used this technique to compute an optical flow for every successive pair of frames of the input sequence independently of all other frames.

2.2. Foreground object segmentation

As discussed in Section 1.2, we would like to analyze the motion patterns of the object of interest. To concentrate on the motion caused by the object only, we need first to obtain the foreground object segmentation map. There are many techniques in object segmentation. However when the background scene of the image sequence is available, background subtraction is an effective and robust technique [9], which is used in this paper. The segmentation map after the initial background subtraction is not in a ready-to-use form, because it contains a considerable amount of "salt and pepper" noise. To clean these up, we use a 5×5-window morphological opening and closing operators to erase small islands and fill the holes. Opening is defined as an erosion followed by a dilation and closing is defined as a dilation followed by an erosion. Compared with plain erosions and dilations, openings and closings have the nice property that they maintain the original boundary of the object being processed (not becoming thinner or fatter) [10]. The noisy background subtraction map and the morphological-filtered map are illustrated in Figure 2 (a) and (b). The small islands from the background objects have been erased and small holes on the foreground object have been filled. A region labeling process gives us the final segmentation map as illustrated in Figure 2 (c).

(a) (b) (c)

Figure 2. Segmentation maps (a)before filtering. (b)after filtering. (c)final map.

2.3. SVD analysis of the spatio-temporal flows

As illustrated in Figure 1, the recovered flow field for each frame is masked by the corresponding foreground object map. All the flow vectors outside the mask are set to zero. In this way we can ensure later analysis is free from irrelevant flows (from background objects).

The original frame-to-frame dense optical flows are both noisy and expensive to analyze. For the flow field of a 320×240-resolution frame, there are $320 \times 240 \times 2 = 153,600$ floating numbers. Even if we only consider the flow field masked by the foreground object map, this number can still be in the order of 50,000. Direct use of the dense flow field is both unreliable and expensive.

In order to obtain a compact representation of the flow fields, we perform SVD over the entire collection of flow fields from a video sequence. To do this, we follow the approach that is by now standard in computer vision and vectorize each flow field. Specifically, we use an approach similar to the one used in [11] and vectorize the flow field using raster-scan order. A flow field (a $2D$ vector at each pixel) containing s pixels is represented as a long 1D vector consisting of $M = 2s$ elements. Since we are interested in segmenting the action of a *moving* person, we translate our flow fields to a moving coordinate system that is centered around the centroid of the segmentation map of the moving person. This way, the flow fields from different pairs of frames are represented in a common objective coordinate system. We concatenate the flow field "vectors" thus obtained from N pairs of frames into an $M \times N$ matrix A.

The matrix A can be decomposed into the following form:

$$A = U\, W\, V^T \tag{1}$$

where U is a $(M \times N)$ column-orthonormal matrix representing the principle component directions; $W = diag(w_1, ..., w_N)$ is a $(N \times N)$ diagonal matrix with positive or zero elements (the singular values) in descending order along the diagonal. These singular values represent the importance (weight) of their corresponding principle components. V is a $(N \times N)$ matrix that encodes the coefficients used to expand A in terms of U. Since the top L, $L < N$, principle components capture a significant amount of information of the original data, it is possible to approximate any column of A by linear combinations of the L most significant principle components:

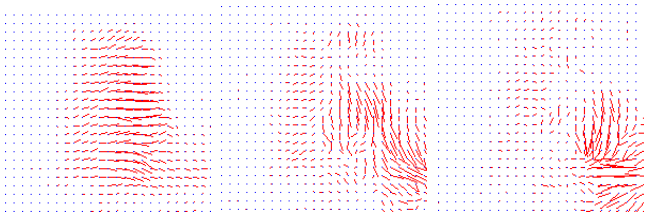
$$\vec{A_n} \approx \sum_{l=1}^{L} \vec{U_l}\, w_l\, v_{nl} \tag{2}$$

where $\vec{A_n}$ denotes the n^{th} column of matrix A (namely the flow-field for the n^{th} frame pair), $\vec{U_l}$ denotes the l^{th} column of matrix U, and v_{nl} is the nl^{th} element of matrix V. After SVD, we have transformed the data from the original pixel-flow space into a flow-basis space. All the analysis discussed in this and later sections will be based on the L-dimensional space spanned by the L most significant SVD component directions.

Figure 3 illustrates the top three flow basis of a flow-field sequence. The basis, especially the top ones, often have physical interpretations. For example, the first basis is a global horizontal translation of the whole human body, basis 2 represents the posture of dropping arms while the whole body moves to the left, and basis 3 corresponds to a significant lower body movement.

2.4. Temporal segmentation of spatial flow patterns

Our hypothesis on detecting visual action boundaries is that the temporal boundaries correspond to *temporal* discontinuity in the *spatial pattern* of image motion. The L SVD coefficients characterize the spatial patterns.



(a) (b) (c)

Figure 3. SVD basis (a)basis 1. (b)basis 2. (c)basis 3

For each of the L SVD coefficients, we obtain a SVD *coefficient evolution curve* by plotting its variation against the frame index n. These curves are the temporal evolutions of the spatial motion patterns. Figures 4 and 5 show the evolution curves of the top three SVD coefficients of two flow-field sequences.

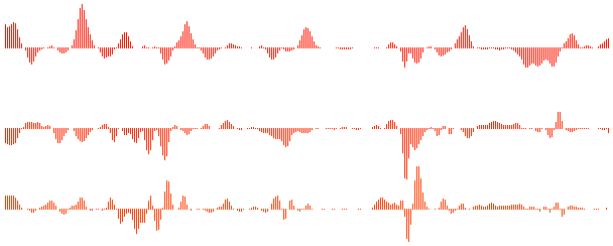


Figure 4. The evolution curves for the top 3 SVD coefficients in the FM sequence.

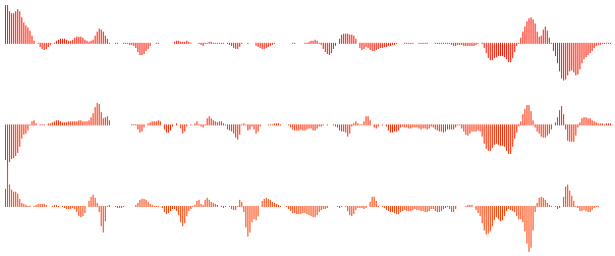


Figure 5. The evolution curves for the top 3 SVD coefficients in the ML sequence.

We next explore three methods to detect segment boundaries based on these evolution curves.

2.4.1 Weighted Euclidean

A flow pattern at a particular time n corresponds to a point in the L-D SVD space. How far away two successive points are in the SVD space signifies if there is a temporal discontinuity in the spatial flow pattern. An obvious approach for detecting the segment boundaries is then to use the pairwise Euclidean distance in the SVD space.

Let d_n be the Euclidean distance between two successive flow-fields (two points in the SVD space). We have

$$d_n = \sum_{l=1}^{L} w_l (v_{nl} - v_{n+1,l})^2 \quad (3)$$

where w_l, the l^{th} diagonal element of matrix W, models each SVD coefficient's weight to the overall distance. After computing d_n for all n's, we obtain a Euclidean distance curve like the one shown in Figure 6.

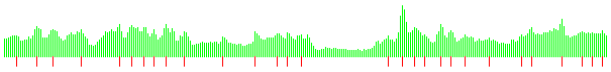


Figure 6. The Euclidean distance curve.

Since the distance measures the degree of temporal discontinuity, we consider the peaks of the curve corresponds to the segment boundaries. The peaks of the curve are detected by using standard local maximum detection techniques, e.g., sign changes of the first derivatives. The vertical bars in the lower portion of Figure 6 show the detected keypose locations.

2.4.2 Linear Prediction

Although the Euclidean distance metric described above is intuitive, in practice we found it to be not very selective in detecting keyposes. This is because the distance measure condenses too much information into a single number and does not discriminate between fine details in the evolution of the coefficients.

Hence, we considered a second method based on linear prediction (LP) error [12]. The logic of this approach is the following: The boundary detection problem can be reformulated as a prediction problem – whenever the prediction error is large, there is a change of pattern thus an action boundary. We next describe how LP technique is used to solve our problem.

For a data sequence $\{y_n, n = 1, ..., N\}$, a future value y_n can be predicted based on existing values [12]:

$$y_n = \sum_{j=1}^{K} p_j y_{n-j} + e_n \quad (4)$$

where K is the number of history data used in prediction, normally a small number such as 5 [12], p_j's are the linear prediction coefficients and e_n is the prediction residue.

Before we can predict y_n, p_j's need to be computed first. To compute p_j's, stationarity (at least piece-wise stationarity) of the data sequence is assumed [12]. That is, the autocorrelation of y_j and y_k depends only on the difference

$|j-k|$, and not on j or k individually. Let Φ_j denote autocorrelation between y_j and y_{n+j}, that is

$$\Phi_j = \frac{1}{N-j} \sum_{n=1}^{N-j} y_n y_{n+j} \quad (5)$$

p_j's can then be computed by the following equations [12]:

$$\begin{aligned} \sum_{j=1}^{K} \Phi_{|j-k|} p_j &= \Phi_k \\ k &= 1, \ldots, K \end{aligned}$$

where there are K equations and K unknowns (p_k). LP has been used extensively in compression and the resulting technique is called linear predictive coding (LPC). In our context, LP provides us a useful tool to detect temporal discontinuities. If there are N flow fields, for each SVD coefficient there is a N-element data sequence. For example, for the l^{th} SVD coefficient, $l = 1, \cdots, L$, there is a data sequence consisting of $\{v_{1l}, \cdots, v_{nl}, \cdots, v_{Nl}\}$. We have L such sequences. Note that U is a column-orthonormal matrix. This orthonormality allows us to use LP for each SVD coefficient sequence independently.

Let E_n be the overall prediction error for flow field n and let E_{nl} be the prediction error along the l^{th} SVD component direction for flow field n. We have

$$\begin{aligned} E_n &= \sum_{l=1}^{L} w_l E_{nl} \\ E_{nl} &= |v_{nl} - \hat{v_{nl}}| \\ \hat{v_{nl}} &= \sum_{j=1}^{K} p_j v_{n-j,l} \end{aligned}$$

where w_l is the weight for coefficient l and $\hat{v_{nl}}$ is the predicted value of v_{nl}. Computing E_n for all the flow fields, we obtain a prediction error curve as shown in Figure 7.

Prediction errors reflect how *discontinuous* a signal is. Therefore the places having large prediction error correspond to the action segment boundaries. By using the same peak-detecting techniques as discussed in Section 2.4.1, we can detect the keypose locations based on the prediction error curve. They are shown as the vertical bars in the lower portion of the figure.

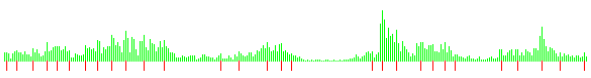


Figure 7. The prediction error curve.

2.4.3 Zerocrossing

The Linear-Prediction error metric is clearly more discriminating than the weighted Euclidean distance metric. However, both metrics treated the SVD evolution curves as ordinary time series and did not take into account the characteristics of these curves. When we visually observed the evolution curves of the coefficients (see Figures 4 and 5), we noticed that SVD coefficients formed groups along the time axis and there are distinct locations where many of the coefficients simultaneously changed signs. These typically correspond to temporal locations where the direction of a particular component of the image motion of the subject has changed. We next develop a detection technique that will take this observation into account.

The group boundaries are marked by coefficient sign changes. Detecting the locations of zerocrossings of the coefficients therefore corresponds to detecting the group boundaries.

Because the evolution curves are noisy, a straightforward utilization of zerocrossing detection will not work. To reduce the noise and to suppress irrelevant sign changes around areas where the coefficient magnitudes are low, we smoothed the evolution curve for each coefficient, and then associated a strength (confidence) with each zerocrossing based on the magnitudes of the evolution curves on both sides of that crossing. We used the areas under the evolution curve as the strength indicator. Let Z_n be the overall zerocrossing strength and Z_{nl} be the zerocrossing strength from SVD coefficient l. We have

$$\begin{aligned} Z_n &= \sum_{l=1}^{L} w_l \, Z_{nl} \\ Z_{nl} &= I(n) * (Ar(n) + Ar(n+1))/2 \end{aligned}$$

where $I(n)$ is an indicator function whose value is one if there is zerocrossing and zero otherwise. $Ar(n)$ is the area under the curves between the previous zerocrossing and the current current zerocrossing and $Ar(n+1)$ is the area between the current and the next zerocrossings. Computing Z_n for all the frames, we obtain a zerocrossing strength curve as shown in Figure 8. Since we conjecture the coefficient group boundaries correspond to, or is a superset of, the action segment boundaries, detecting action boundaries therefore corresponds to detecting peaks of the zerocrossing strength curve. The vertical bars in the lower portion of the figure are the detected group boundaries.

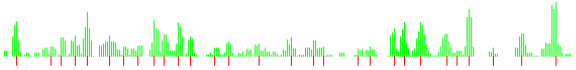


Figure 8. The zerocrossing strength curve.

3. Experimental Results

3.1. Data sets

We conducted experiments with several video sequences of people performing common household chores and en-

gaging in other such activities. In this section, we describe the results of our techniques on two sequences showing a person making a bed. These sequences were obtained from Zacks who used these in his psychological study [6]. One sequence, which we refer to as the *ML sequence* shows a male subject making a bed, while the other, which we refer to as the *FM sequence* shows a female subject making a bed. The original video sequences were recorded on VHS tapes. We digitized the video sequences into MPEG compressed files at a frame-rate of 10 frames/second. The resolution of each frame is 320×240 pixels. Each video sequence has 300 frames.

3.2. Keypose detection results

Due to page limitation, in Figures 9 and 10 we only show the keyposes detected by using the zerocrossing method. In all the experiments reported in this section, $L = 10$, $K = 5$, and γ = 100%, where γ is the ratio between the number of detected boundaries and the number of psychological study boundaries as defined in Section 3.4.

The detected keyposes closely match our visual examination – 39 out of 42 detected boundaries are the frames that the subject is changing actions. For example, in Figure 9, keypose 1 is the place when the subject stops and begins to pick the blanket. Keypose 2 is the place when the subject pulls the blanket to the highest point and starts to put it down. In Figure 10, keypose 1 is where the subject starts to turn her body from left to right. Keypose 2 is where the subject begins to fold the blanket using her arms. Among successive pair of keyposes, the more similar they are in motion patterns, the lower the zerocrossing strength is for one of them. For examples, in Figure 10, keypose 8's motion pattern is similar to that of keypose 9. If we look at this sequence's zerocrossing strength curve (see Figure 8), keypose 8 has very limited zerocrossing strength. This matches our intuition closely.

3.3. Quantitative evaluation

A visual examination of the results indicates that the keypose frames detected by our methods match the locations where there is a transition from one smooth body movement to another of the subject. However, a more "objective" evaluation of our results is desirable. The authors of the psychological studies [6], Zacks, *et al.*, have kindly provided us with both the video sequences and the human segmented boundaries. These provide us with a possible source for a more "objective" analysis. In this section, we describe our preliminary results on making this comparison.

In the course of their research, they asked 16 human observers to independently mark the action boundaries in input video sequences, including the two that we used for our experiments. The observers were asked to mark the boundaries of the "smallest" meaningful units of action seen in the video. Our expectation is that this "fine-grained" segmentation by human observers is likely to be correlated with the automatic keyposes detected by our methods. The human segmentations we obtained from the authors are in 0.1-second resolution, which corresponds to 1-frame resolution in our MPEG digitized video sequences.

3.4. Evaluation methodology

We used the boundary detection data from the 16 subjects to create a histogram of boundary detections as a function of frame index. We interpret the peaks of these histograms correspond to possible segment boundaries. Since the human observers are looking for the "smallest *meaningful* chunks", we expect the human detected boundaries to be a subset of our *lower-level* segmentation. To measure the correlation between the human segmentation and our detected boundaries, we devised the following approach:

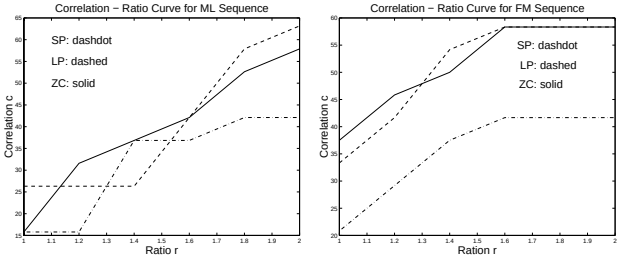
Let N_p be the total number of boundaries in the psychological study (whose locations are given by the histogram peaks), N_{pd} be the number of psychological study boundaries that have also been detected by our proposed methods, and N_d be the total number of boundaries detected by our methods. We define the following measure c as the measure of correlation between the psychological study data and our results:

$$c = \frac{N_{pd}}{N_p} \times 100\% \tag{6}$$

Our detection methods involve thresholds that determine if a local maximum is significant. This gives a degree of freedom to choose the number of detections N_d. Since we expect that our methods will detect more boundaries than those from the psychological study, we can make the comparison for different choices of N_d . Define γ to be the ratio between N_d and N_p, i.e.,

$$\gamma = \frac{N_d}{N_p} \times 100\% \tag{7}$$

We varied the value of γ to see how it affects the correlation between our data and those from the psychological study.



(a) (b)

Figure 11. Correlation vs. γ (a)ML sequence. (b)FM sequence.

Figure 9. Keyposes detected by ZC for the ML sequence when $\gamma = 100\%$.

3.5. Results and observations

Let SP denote the simple Euclidean method, LP denote the linear prediction method, and ZC denote the zerocrossing method. Figure 11 shows the experimental results for the ML and FL sequences and how the correlation c changes with γ, where γ ranges from 100% to 200% with a 20% increment. The dashdot curves are for SP, dashed curves for LP, and solid curves for ZC. Based on the figure, the following observations can be made:

- As expected, the SP method does not perform well. Its correlation with the psychological experiments is consistently lower than that of the other two methods.
- The performances of LP and ZC are comparable, with LP slightly better in the ML sequence and ZC better in the FM sequence. By looking at the evolution curves (Figures 4 and 5) it is clear that ML sequence has less distinguished groups in the curves than the FM sequence. Since ZC is designed to detect the groups in evolutions curves and LP is independent of the characteristics of the curves, ZC performs better in the FM sequence while LP performs better in the ML sequence.
- Our segmentation technique is a bottom-up approach based on *spatio-temporal* motion patterns only. Around 60% correlation at $\gamma = 180\%$ (by ZC and LP) with the human annotated boundaries is a very encouraging result.

4. Conclusions

In this paper, we have investigated the problem of segmenting a continuous video sequence containing human actions into shorter subsequences corresponding to smooth or continuous visual action segments. Specifically, we detect *keypose* frames which correspond to locations of *temporal* discontinuities in the *spatial* pattern of image motion. While the literature on visual action and activity analysis has focused heavily on recognition and to a lesser extent on activity *detection* for a known and limited set of action "models", the problem of segmenting a general unrestricted video sequence containing human action has not been previously addressed.

We present a complete algorithm that takes input video sequences, computes frame to frame optical flow, projects the flow-fields into a basis set using SVD analysis, and detects temporal discontinuities in the trajectories of the basis coefficients over time. The keyposes correspond to frames at these discontinuities. We have investigated three possible measures for detecting keyposes in this fashion.

We present results on two real video sequences of human subjects performing common household activities. A visual examination of the keyposes detected by our methods indicate that they correspond to natural locations of discontinuities. We also measured the correlation between our segmentation results and data that we obtained from an independent psychological study on human observers segmenting the same sequences. Our preliminary results are very encouraging as shown in Section 3.

There are a number of potential applications of the type of segmentation described here. First, the keyposes serve as a compact visual summary of the actions. Second, these low-levels segment boundaries are useful building blocks for higher level action segmentation and recognition. Finally, the segment boundaries detected by our technique can also serve as better locations for I-frames in an MPEG like

Figure 10. Keyposes detected by ZC for the FL sequence when $\gamma = 100\%$.

compression scheme.

References

[1] M. Black, Y. Yacoob, and X. S. Ju, *Recognizing human motion using parameterized models of optical flow*. Boston, MA: Kluwer Academic Publishers, 1997.

[2] Y. Yacoob and M. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232–247, 1999.

[3] J. Davis and A. Bobick, "The representation and recognition of action using temporal templates," in *Proc. IEEE Conf. on Comput. Vis. and Patt. Recog.*, July 1997.

[4] J. L. Mitchell, D. L. Gall, and C. Fogg, *MPEG Video Compression Standard*. Champman and Hall, 1996.

[5] U. Gargi, R. Ksturi, and S. Antani, "Performance characterization and comparison of video indexing algorithms," in *Proc. IEEE Conf. on Comput. Vis. and Patt. Recog.*, pp. 559–565, 1998.

[6] J. Zacks, B. Tversky, and G. Iyer, "Perceiving, remembering, and communicating structure in events," *Journal of Experimental Psychology: General (in press)*, 2000.

[7] D. Newtson, "Attribution and the unit of perception of ongoing behavior," *Journal of Personality and Social Psychology*, vol. 28, pp. 28–38, 1973.

[8] R. Szeliski, "A multi-view approach to motion and stereo," in *Proc. IEEE Conf. on Comput. Vis. and Patt. Recog.*, July 1999.

[9] K. Yoyama, J. Krumm, B. Brumitt, and B. Meyes, "Wallflower: Principles and practice of background maintenance," in *Proc. IEEE Int. Conf. on Comput. Vis.*, 1999.

[10] A. K. Jain, *Fundamentals of Digital Image processing*. Prentice Hall.

[11] M. Black, Y. Yacoob, A. Jepson, and D. Fleet, "Learning parameterized models of image motion," in *Proc. IEEE Conf. on Comput. Vis. and Patt. Recog.*, June 1997.

[12] J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Macmillan Publishing Company, 1992.